DravidianLangTech 2025

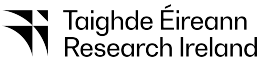
The Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages

Proceedings of the Workshop

May 3, 2025

The DravidianLangTech organizers gratefully acknowledge the support from the following sponsors.

In cooperation with



©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-228-2

Introduction

We are excited to welcome you to the fifth workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech 2025), the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025). This year, the workshop will be held in a hybrid format (both online and at Acoma, The Albuquerque Convention Centre, Albuquerque, New Mexico) on May 3rd, 2025, alongside NAACL 2025, which will take place from April 29 to May 4, 2025. With the rapid advancement of technology, digital communication has become a central part of daily life. While many globally dominant languages have successfully transitioned into the digital era, numerous regional and low-resource languages continue to face significant technological challenges. A prominent example of such a language family is the Dravidian language family, which remains underrepresented in the domains of speech and natural language processing (NLP). The DravidianLangTech-2025 workshop series aims to address the technological marginalization of Dravidian languages by fostering research and development in computational linguistics, speech processing, and NLP specific to these languages. By building inclusive language technologies, the goal is to ensure equitable access to digital information and communication tools for monolingual speakers of Dravidian languages. These workshops represent an important step toward preserving linguistic diversity and preventing the digital extinction of these historically and culturally significant languages. This will be the fifth workshop on speech and language technologies for Dravidian languages, continuing our mission to advance technological solutions and promote linguistic inclusivity. The workshop received a total of 204 active submissions. Reviewer recruitment was highly successful with 405 reviewers accepting invitations. Of the 1,371 assigned reviews, 758 were completed, achieving a review submission rate of 55.29%. Additionally, 49.39% of reviewers (204 out of 413) fulfilled all their assigned tasks. Among the submissions, 81.86% (167 out of 204) received at least three reviews, reflecting a comprehensive evaluation process. Decisions were made for all submissions (100%), resulting in an overall acceptance rate of 57.84% (120 papers). This included 10 papers (4.90%) accepted for oral presentations and 118 papers (57.84%) accepted for poster presentations. The remaining 76 papers (37.25%) were rejected after review. These statistics highlight a rigorous yet inclusive selection process supported by dedicated reviewers and a commitment to academic excellence.

Program Committee

Program Chairs

Bharathi Raja Chakravarthi, University of Galway, Ireland
Ruba Priyadharshini, The Gandhigram Rural Institute - Deemed University, Tamil Nadu, India
Anand Kumar Madasamy, National Institute of Technology Karnataka, India
Sajeetha Thavareesan, Eastern University, Sri Lanka
Elizabeth Sherly, Digital University Kerala, India
Saranya Rajiakodi, Central University of Tamil Nadu, India
Balasubramanian Palani, Indian Institute of Information Technology (IIIT) Kottayam, India
Malliga Subramanian, Kongu Engineering College, Tamil Nadu, India
Dhivya Chinnappa, USA
Subalalitha Chinnaudayar Navaneethakrishnan, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Publication Chairs

Rahul Ponnusamy, Data Science Institute, University of Galway, Ireland
Shunmuga Priya Muthusamy Chinnan, Data Science Institute, University of Galway, Ireland
Prasanna Kumar Kumaresan, Data Science Institute, University of Galway, Ireland

Programme Committee

A. Justin Gopinath, Vellore Institute of Technology, India
Aathavan Nithiyananthan, University of Moratuwa, Sri Lanka
Abdullah Al Nahian, American International University, Bangladesh
Abdur Rahman, Shahjalal University of Science and Technology, Bangladesh
Abhai Pratap Singh, Carnegie Mellon University, USA
Abhay Vishwakarma, Motilal Nehru National Institute Of Technology, India
Abhinav Kumar, Sikha 'O' Anushandhan University, India
Abirami Jayaraman, Sri Sivasubramaniya Nadar Institutions, India
Adarsh Valoor, University of Southampton, UK
Adeep Hande, Comcast Applied AI, USA
Advait Vats, Indian Institute of Technology, Madras, India
Ahamed Rameez Mohamed Nizzad, British College of Applied Studies, Sri Lanka
Aishwarya Gaddam, George Mason University, USA
Aishwarya Selvamurugan, Sri Eshwar College of Engineering, India
Alessandra Teresa Cignarella, Ghent University, Belgium
Amit Jaspal, Facebook, USA
Angel Deborah S, Sri Sivasubramaniya Nadar College of Engineering, India
Anik Mahmud Shanto, Chittagong University of Engineering and Technology, Bangladesh
Anindo barua, Chittagong University of Engineering and Technology, Bangladesh
Anisha Ahmed, Shahjalal University of Science and Technology, Bangladesh
Anusha M D Gowda, Yenepoya Deemed to be University, India
Aravindh M, RMK Engineering College, India
Arivuchudar K, RMK Engineering College, India
Arpita Mallik, Chittagong University of Engineering and Technology, Bangladesh
Arthi R, RMK Engineering College, India

Arun Prasad T D, Amrita Vishwa Vidyapeetham (Deemed University), India
Aruna Devi Shanmugam, Sri Sivasubramaniya Nadar Institutions, India
Arunaggiri Pandian Karunanidhi, Micron Technology, USA
Arupa Barua, Chittagong University of Engineering and Technology, Bangladesh
Arya Palackal Shijish, Amrita Vishwa Vidyapeetham (Deemed University), India
Asha Hegde, Mangalore University, India
Ashim Dey, Chittagong University of Engineering and Technology, Bangladesh
Ashok Yadav, Indian Institute of Information Technology Allahabad, India
Ashraf Deen, RMK Engineering College, India
Ashraful Islam Paran, Chittagong University of Engineering and Technology, Bangladesh
Ashutosh Tripathi, Rajiv Gandhi Institute of Petroleum Technology, India
Avaneesh Koushik, Sri Sivasubramaniya Nadar Institutions, India
Azmine Toushik Wasi, Shahjalal University of Science and Technology, Bangladesh
B Saathvik, Sri Sivasubramaniya Nadar Institutions, India
Babatunde Abimbola Abiola, Cape Peninsula University of Technology, South Africa
Bachu Naga sri Harini, Amrita Vishwa Vidyapeetham (Deemed University), India
Bagavathi C, Amrita Vishwa Vidyapeetham (Deemed University), India
Bharathi B, Sri Sivasubramaniya Nadar College Of Engineering, India
Bhuvaneswari Sivagnanam, Central University of Tamil Nadu, India
Billodal Roy, Lowe's, USA
Bitan Mallik, Vellore Institute of Technology, India
Bommineni Sahitya, RMK Engineering College, India
Boomika E, RMK Engineering College, India
Burugu Rahul, Amrita Vishwa Vidyapeetham (Deemed University), India
Dhanyashree G, RMK Engineering College, India
Dharunika Sasikumar, Sri Sivasubramaniya Nadar Institutions, India
Diya Seshan, Sri Sivasubramaniya Nadar Institutions, India
Dola Chakraborty, Chittagong University of Engineering and Technology, Bangladesh
Dondluru Keerthana, Amrita Vishwa Vidyapeetham (Deemed University), India
Durga Prasad Manukonda, ASRlytics, India
Enjamamul Haque Eram, Shahjalal University of Science and Technology, Bangladesh
Eric SanJuan, Université d'Avignon, France
Eshwanth Karti T R, Amrita Vishwa Vidyapeetham (Deemed University), India
Fariha Haq, Chittagong University of Engineering and Technology, Bangladesh
Farjana Alam Tofa, Chittagong University of Engineering and Technology, Bangladesh
Fatima Uroosa, Instituto Politécnico Nacional, Mexico
Fiona Victoria Stanley Jothiraj, Oregon State University, USA
Fred Philippy, University of Luxemburg, Luxemburg
G Manikandan, RMK Engineering College, India
Ganesh Sundhar S, Amrita Vishwa Vidyapeetham, India
Gersome Shimi, Madras Christian College, India
Girma Yohannis Bade, Instituto Politécnico Nacional, Mexico
Gladiss Merlin N.R, RMK Engineering College, India
Habiba A, National Institute of Technology Puducherry, India
Hamada Nayel, Prince Sattam bin Abdulaziz University, Saudi Arabia
Hare Ram C, RMK Engineering College, India
Hari Krishnan N, Amrita Vishwa Vidyapeetham (Deemed University), India
Harish Vijay V, Amrita Vishwa Vidyapeetham (Deemed University), India
Harry Thomas Kandathil, Vellore Institute of Technology, India
Harshita Sharma, University of Delhi, India
Hasan Murad, Chittagong University of Engineering and Technology, Bangladesh

Hosahalli Lakshmaiah Shashirekha, Mangalore University, India
Ippatapu Venkata Srichandra, Amrita Vishwa Vidyapeetham (Deemed University), India
J Bhuvana, Sri Sivasubramaniya Nadar College of Engineering, India
Jahnavi Murali, Sri Sivasubramaniya Nadar Institutions, India
Janeshvar Sivakumar, Sri Sivasubramaniya Nadar Institutions, India
Jannath Nisha O S, Vellore Institute of Technology, India
Jerin Mahibha C, Meenakshi Sundararajan Engineering College, India
Jobin Jose, Indian Institute of Information Technology Kottayam, India
Jubeerathan Thevakumar, University of Moratuwa, Sri Lanka
Jyothish Lal G, Amrita Vishwa Vidyapeetham (Deemed University), India
K Anishka, Sri Sivasubramaniya Nadar College Of Engineering, India
Kalaivani K S, Kongu Engineering College, India
Kalpana K, RMK Engineering College, India
Kankipati Venkata Meghana, Amrita Vishwa Vidyapeetham (Deemed University), India
Kasu Sai Kartheek Reddy, Indian Institute of Information Technology Dharwad, India
Kavin Bharathi, Vellore Institute of Technology, India
Kawsar Ahmed, Chittagong University of Engineering and Technology, Bangladesh
Keerthana NNL, Indian Institute of Information Technology Kottayam, India
Keerthi Vasan A, RMK Engineering College, India
Khadiza Sultana Sayma, Chittagong University of Engineering and Technology, Bangladesh
Kogilavani Shanmugavadivel, kongu engineering college, India
Kondakindi Supriya, Amrita Vishwa Vidyapeetham (Deemed University), India
Konkimalla Laxmi Vignesh, Amrita Vishwa Vidyapeetham (Deemed University), India
Kritika A, Amrita Vishwa Vidyapeetham (Deemed University), India
Lahari P, RMK Engineering College, India
Lalith Kishore V P, RMK Engineering College, India
Lekhashree A, RMK Engineering College, India
Lemlem Eyob Kawo, Instituto Politécnico Nacional, Mexico
Livin Nector Dhasan, Indian Institute of Technology, Madras, India
Luxshan Thavarasa, University of Moratuwa, Sri Lanka
Mahankali Sri Ram Krishna, Amrita Vishwa Vidyapeetham (Deemed University), India
Mahfuz Ahmed Anik, Shahjalal University of Science and Technology, Bangladesh
Mahir Absar Khan, Shahjalal University of Science and Technology, Bangladesh
Manan Buddhadev, Rochester Institute of Technology, USA
Marc Brysbaert, Universiteit Gent, Belgium
Md Ayon Mia, Dhaka International University, Bangladesh
Md Minhazul Kabir, Chittagong University of Engineering and Technology, Bangladesh
Md Mizanur Rahman, Chittagong University of Engineering and Technology, Bangladesh
MD Musa Kalimullah Ratul, Chittagong University of Engineering and Technology, Bangladesh
Md. Mohiuddin, Chittagong University of Engineering and Technology, Bangladesh
Md. Alam Miah, Chittagong University of Engineering and Technology, Bangladesh
Md. Mahadi Rahman, Chittagong University of Engineering and Technology, Bangladesh
Md. Refaj Hossan, Chittagong University of Engineering and Technology, Bangladesh
Md. Sajid Alam Chowdhury, Chittagong University of Engineering and Technology, Bangladesh
Md. Sajjad Hossain, Chittagong University of Engineering and Technology, Bangladesh
Md. Tanvir Ahammed Shawon, Chittagong University of Engineering and Technology, Bangladesh
Md. Zahid Hasan, Chittagong University of Engineering and Technology, Bangladesh
Meenakshy S, Amrita Vishwa Vidyapeetham (Deemed University), India
Meetesh Saini, Vellore Institute of Technology, India
Mesay Gemeda Yigezu, Instituto Politécnico Nacional, Mexico
Mikhail Krasitskii, Instituto Politécnico Nacional, Mexico

Minhaz Chowdhury, Shahjalal University of Science and Technology, Bangladesh
Minoru Sasaki, Ibaraki University, Japan
Miriam Butt, Universität Konstanz, Germany
Mohan Raj, Monash University, Malaysia
Mohan Raj M A, RMK Engineering College, India
Momtazul Arefin Labib, Chittagong University of Engineering and Technology, Bangladesh
Monorama Swain, Copenhagen University, Denmark
Moogambigai A, Sri Sivasubramaniya Nadar College Of Engineering, India
Mugilkrishna D U, Sri Sivasubramaniya Nadar College Of Engineering, India
N.Nasurudeen Ahamed, United Arab Emirates University, UAE
Naihao Deng, University of Michigan - Ann Arbor, USA
Navneet Agarwal, Université de Caen Basse Normandie, France
Navneet Krishna Chukka, Amrita Vishwa Vidyapeetham (Deemed University), India
Nazmus Sakib, Chittagong University of Engineering and Technology, Bangladesh
Nida Hafeez, Instituto Politecnico Nacional, Mexico
Niranjan kumar M, Lowe's, USA
Nishanth S, Amrita Vishwa Vidyapeetham (Deemed University), India
Nithish Ariyha K, Amrita Vishwa Vidyapeetham (Deemed University), India
Nitisha Aggarwal, University of Delhi, India
Noor Mairukh Khan Arnob, University of Asia Pacific, Bangladesh
Olga Kolesnikova, Instituto Politécnico Nacional, Mexico
Pandiarajan D, Sri Sivasubramaniya Nadar Institutions, India
Pathange Omkareshwara Rao, Amrita Vishwa Vidyapeetham (Deemed University), India
Pavithra J, RMK Engineering College, India
Ponsubash Raj R, Sri Sivasubramaniya Nadar College Of Engineering, India
Poojitha Sai Manikandan, Amrita Vishwa Vidyapeetham (Deemed University), India
Pranav Gupta, Lowes Inc
Premjith B, Amrita Vishwa Vidyapeetham (Deemed University), India
Priyatharshan Balachandran, University of Moratuwa, Sri Lanka
Rahatun Nesa Priti, Shahjalal University of Science and Technology, Bangladesh
Raj Sonani, Cornell University, USA
Rajalakshmi Sivanaiah, Sri Sivasubramaniya Nadar College of Engineering, India
Raksha Adyanthaya, Yenepoya Institute of Arts, Science, Commerce and Management, India
Ramesh Kannan R, Vellore Institute of Technology, India
Rasha Sharma, Amrita Vishwa Vidyapeetham (Deemed University), India
Ratnavel Rajalakshmi, Vellore Institute of Technology, India
Ravi Teja Potla, NVIDIA, USA
Riya Rajeev, Amrita Vishwa Vidyapeetham (Deemed University), India
Rohan R, Sri Sivasubramaniya Nadar Institutions, India
Rohith Gowtham Kodali, asrlytics, India
S Ananthasivan, Amrita Vishwa Vidyapeetham (Deemed University), India
Sabik Aftahee, Chittagong University of Engineering and Technology, Bangladesh
Sabrina Afroz Mitu, Shahjalal University of Science and Technology, Bangladesh
Safiul Alam Sarker, Chittagong University of Engineering and Technology, Bangladesh
Sai Koneru, Pennsylvania State University, USA
Sakkthi Gurru D, Vellore Institute of Technology, India
Sandra Johnson, RMK Engineering College, India
Sangeetha S, Kumaraguru College of Technology, India
Santhosh Kakarla, George Mason University, USA
Sarbajeet Pattanaik, Indian Institute of Information Technology Allahabad, India
Sarumathi P, Sri Sivasubramaniya Nadar College Of Engineering, India

Satya Subrahmanya Gautama Shastry Bulusu Venkata, George Mason University, USA
Saurabh Aggarwal, Autodesk, USA
Shamima Afroz, Chittagong University of Engineering and Technology, Bangladesh
Shankari S R, Sri Sivasubramaniya Nadar College Of Engineering, India
Shanmitha Thirumoorthy, Vellore Institute of Technology, India
Sharon Sunny, University of Southampton, UK
Sheikh Ayatur Rahman, BRAC University, Bangladesh
Shreyas Karthik, Sri Sivasubramaniya Nadar Institutions, India
Shriya Alladi, Sri Sivasubramaniya Nadar Institutions, India
Shruthi Rengarajan, Amrita Vishwa Vidyapeetham (Deemed University), India
Shruthikaa V, Amrita Vishwa Vidyapeetham (Deemed University), India
Shuang Ao, University of Southampton, UK
Siddhaarth Sekar, Amrita Vishwa Vidyapeetham (Deemed University), India
Sidney Wong, University of Canterbury, New Zealand
Simran , Institute of Informatics and Communication, India
Sivasuthan Sukumar, University of Moratuwa, Sri Lanka
Somsubhra De, Indian Institute of Technology, Madras, India
Sourabh Deoghare, Indian Institute of Technology, Bombay
Souvik Bhattacharyya, Lowe's, USA
Sowmya Vajjala, National Research Council Canada, Canada
Spurthi Amba Hombaiah, Google DeepMind, USA
Sreeja K, Sri Sivasubramaniya Nadar College of Engineering, India
Srihari V K, Sri Sivasubramaniya Nadar Institutions, India
Srijita Dhar, Chittagong University of Engineering and Technology, Bangladesh
Sripriya N, Sri Sivasubramaniya Nadar College of Engineering, India
Subhashini Sudhakar, Amrita Vishwa Vidyapeetham (Deemed University), India
Suriya KP, Amrita Vishwa Vidyapeetham (Deemed University), India
Swetha N.G, Vellore Institute of Technology, India
Syeda Alisha Noor, Chittagong University of Engineering and Technology, Bangladesh
Symom Hossain Shohan, Chittagong University of Engineering and Technology, Bangladesh
Tanisha Sriram, Sri Sivasubramaniya Nadar College Of Engineering, India
Tara Samiksha, Amrita Vishwa Vidyapeetham (Deemed University), India
Tareque Md Hanif, Chittagong University of Engineering and Technology, Bangladesh
Temitope Oladepo, Federal University Oye-Ekiti, Nigeria
Tewodros Achamaleh, Instituto Politécnico Nacional, Mexico
Thenmozhi Durairaj, SSN College of Engineering, India
Tofayel Ahmmed Babu, Chittagong University of Engineering and Technology, Bangladesh
Tolulope Olalekan Abiola, Instituto Politécnico Nacional, Mexico
Trina Chakraborty, Shahjalal University of Science and Technology, Bangladesh
Udoy Das, East Delta University, Bangladesh
Uma Jothi, Amrita Vishwa Vidyapeetham (Deemed University), India
V Guru charan, Collaborative Dynamics, USA
Vajratiya Vajrobol, University of Delhi, India
Venkatesh Velugubantla, Meridian Cooperative, USA
Vijay Karthick Vaidyanathan, Sri Sivasubramaniya Nadar Institutions, India
Vijay Manickam R, Vellore Institute of Technology, India
Vikash J, Amrita Vishwa Vidyapeetham (Deemed University), India
Vishal A S, Amrita Vishwa Vidyapeetham (Deemed University), India
Vrijendra Singh, Indian Institute of Information Technology Allahabad, India
Wahid Faisal, Shahjalal University of Science and Technology, Bangladesh
Yeshwanth Balaji A P, Amrita Vishwa Vidyapeetham (Deemed University), India

Yves Scherrer, University of Oslo, Norway

Best Reviewers

Ashutosh Tripathi, Rajiv Gandhi Institute of Petroleum Technology, India
Adeep Hande, Comcast Applied AI, USA
Amit Jaspal, Facebook, USA

Active Reviewers

Md. Refaj Hossan, Chittagong University of Engineering and Technology, Bangladesh
Mahir Absar Khan, Shahjalal University of Science and Technology, Bangladesh
Bitan Mallik, Vellore Institute of Technology, India
Durga Prasad Manukonda, ASRlytics, India
Jannath Nisha O S, Vellore Institute of Technology, India
Arunaggiri Pandian Karunanidhi, Micron Technology, USA
Fred Philippy, University of Luxemburg, Luxemburg
B Saathvik, Sri Sivasubramaniya Nadar Institutions, India
Sowmya Vajjala, National Research Council Canada, Canada
Ashok Yadav, Indian Institute of Information Technology Allahabad, India

Keynote Talk: Understanding Attention in Asymmetric Kernel Point of View

Dr. Soman K. P.
Amrita Vishwa Vidyapeetham, India
2025-05-03 09:15 – Room: **Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico, USA**

Abstract: Transformers has redefined deep learning research and has become the most prominent architecture across domains such as natural language processing, computer vision, and image processing. Attention mechanism, particularly self-attention, is central to the success of this architecture, which allows the model to capture dependencies across the input sequences. However, the fundamental challenge in understanding self-attention is its intrinsic symmetry. The existing works often consider self-attention as a kernel method, leveraging symmetric kernels based on Mercer's theorem. However, the self-attention matrices used in the transformer architectures are inherently asymmetric, which leads to an inconsistency between the theoretical formulation and the practical implementation. The primal-attention, a novel attention mechanism based on kernel singular value decomposition explicitly models the asymmetry. Therefore, reformulating self-attention using primal-dual representation ensures efficient computation and low-rank approximation that enhances performance and generalization.

Bio: Dr. Soman K. P. is the Dean of the School of Artificial Intelligence and Head of the Department at Amrita Vishwa Vidyapeetham, Coimbatore. With over 27 years of experience in research and teaching, his expertise spans Artificial Intelligence and Data Science. He has published more than 500 papers in leading journals and conferences, including IEEE Transactions, IEEE Access, and Applied Energy. He is the author of four books, including Insight into Wavelets, Insight into Data Mining (also translated into Chinese), Support Vector Machines and Other Kernel Methods, and Signal and Image Processing—the Sparse Way. Dr. Soman is the most cited researcher with over 10,000 citations. He has consistently been ranked among the world's top 2% most influential scientists by Stanford University for the past three years. His contributions have also been recognized by the Government of India and organizations like Springer Nature and Career 360. At CEN, he leads M.Tech programs in Computational Engineering and Networking (Data Science) and Computer Science and Engineering (Artificial Intelligence). A new B.Tech program in AI and Data Science launched under his leadership in 2023. He has guided over 20 Ph.D. scholars and currently supervises 8+ ongoing doctoral researchers. His current research interests include AI for DNA sequence analysis, reinforcement learning in robotics, computer vision, and cyber-physical systems.

Table of Contents

Program

Saturday, May 3, 2025

09:00 - 09:15 ***Opening Remarks***

09:15 - 09:45 ***Understanding Attention in Asymmetric Kernel Point of View***

09:45 - 10:30 ***Oral Session 1***

09:45 - 10:00 *F^2 (FutureFiction): Detection of Fake News on Futuristic Technology*
Msvpj Sathvik, Venkatesh Velugubantla and Ravi Teja Potla

10:00 - 10:15 *TSD: Towards Computational Processing of Tamil Similes - A Tamil Simile Dataset*
Aathavan Nithiyananthan, Jathushan Raveendra and Uthayasanker Thayasivam

10:15 - 10:30 *Towards Effective Emotion Analysis in Low-Resource Tamil Texts*
Priyatharshan Balachandran, Uthayasanker Thayasivam, Randil Pushpananda and Ruvan Weerasinghe

10:30 - 11:00 ***Tea Break***

11:00 - 12:30 ***Oral Session 2***

11:00 - 11:15 *Bridging Linguistic Complexity: Sentiment Analysis of Tamil Code-Mixed Text Using Meta-Model*
Anusha M D Gowda, Deepthi Vikram and Parameshwar R Hegde

11:15 - 11:30 *Misogynistic Meme Detection in Dravidian Languages Using Kolmogorov Arnold-based Networks*
Manasha Arunachalam, Navneet Krishna Chukka, Harish Vijay V, Premjith B and Bharathi Raja Chakravarthi

11:30 - 11:45 *Detection of Religious Hate Speech During Elections in Karnataka*
Msvpj Sathvik, Raj Sonani and Ravi Teja Potla

11:45 - 12:00 *DravLingua@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages using Late Fusion of Muril and Wav2Vec Models*
Aishwarya Selvamurugan

12:00 - 12:15 *Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian languages: DravidianLangTech@NAACL 2025*
Jyothish Lal G, Premjith B, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan and Ratnavel Rajalakshmi

Saturday, May 3, 2025 (continued)

12:15 - 12:30 *Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025*
Premjith B, Nandhini Kumaresh, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, Sajeetha Thavareesan and Prasanna Kumar Kumaresan

12:30 - 14:15 ***Lunch Break***

14:15 - 15:30 ***Oral Session 3***

14:15 - 14:30 *Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025*
Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Raja Meenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagan Jananayagan and Kishore Kumar Ponnusamy

14:30 - 14:45 *Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025*
Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam and Anshid K A

14:45 - 15:00 *Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu*
Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hegde, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Krishnakumari K, Charmathi Rajkumar, Poorvi Shetty and Harshitha S Kumar

15:00 - 15:15 *Overview on Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments: DravidianLangTech@NAACL 2025*
Bharathi Raja Chakravarthi, Saranya Rajiakodi, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi and Rohan R

15:15 - 15:30 *Overview of the Shared Task on Fake News Detection in Dravidian Languages-DravidianLangTech@NAACL 2025*
Malliga Subramanian, Premjith B, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Balasubramanian Palani and Bharathi Raja Chakravarthi

15:30 - 16:00 ***Tea Break***

16:00 - 17:30 ***Poster Session***

16:00 - 17:30 *Incepto@DravidianLangTech 2025: Detecting Abusive Tamil and Malayalam Text Targeting Women on YouTube*
Luxshan Thavarasa, Sivasuthan Sukumar and Jubeerathan Thevakumar

16:00 - 17:30 *Eureka-CIOL@DravidianLangTech 2025: Using Customized BERTs for Sentiment Analysis of Tamil Political Comments*
Enjamamul Haque Eram, Anisha Ahmed, Sabrina Afroz Mitu and Azmine Toushik Wasi

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *Akatsuki-CIOL@DravidianLangTech 2025: Ensemble-Based Approach Using Pre-Trained Models for Fake News Detection in Dravidian Languages*
Mahfuz Ahmed Anik, Md. Iqramul Hoque, Wahid Faisal, Azmine Toushik Wasi and Md Manjurul Ahsan

16:00 - 17:30 *RMKMavericks@DravidianLangTech 2025: Tackling Abusive Tamil and Malayalam Text Targeting Women: A Linguistic Approach*
Sandra Johnson, Boomika E and Lahari P

16:00 - 17:30 *RMKMavericks@DravidianLangTech 2025: Emotion Mining in Tamil and Tulu Code-Mixed Text: Challenges and Insights*
Gladiss Merlin N.r, Boomika E and Lahari P

16:00 - 17:30 *JAS@DravidianLangTech 2025: Abusive Tamil Text targeting Women on Social Media*
B Saathvik, Janeshvar Sivakumar and Thenmozhi Durairaj

16:00 - 17:30 *Team-Risers@DravidianLangTech 2025: AI-Generated Product Review Detection in Dravidian Languages Using Transformer-Based Embeddings*
Sai Sathvik, Muralidhar Palli, Keerthana Nnl, Balasubramanian Palani, Jobin Jose and Siranjeevi Rajamanickam

16:00 - 17:30 *NLPopsCIOL@DravidianLangTech 2025: Classification of Abusive Tamil and Malayalam Text Targeting Women Using Pre-trained Models*
Abdullah Al Nahian, Mst Rafia Islam, Azmine Toushik Wasi and Md Manjurul Ahsan

16:00 - 17:30 *AiMNLP@DravidianLangTech 2025: Unmask It! AI-Generated Product Review Detection in Dravidian Languages*
Somsubhra De and Advait Vats

16:00 - 17:30 *byteSizedLLM@DravidianLangTech 2025: Fake News Detection in Dravidian Languages Using Transliteration-Aware XLM-RoBERTa and Transformer Encoder-Decoder*
Durga Prasad Manukonda and Rohith Gowtham Kodali

16:00 - 17:30 *byteSizedLLM@DravidianLangTech 2025: Fake News Detection in Dravidian Languages Using Transliteration-Aware XLM-RoBERTa and Attention-BiLSTM*
Rohith Gowtham Kodali and Durga Prasad Manukonda

16:00 - 17:30 *byteSizedLLM@DravidianLangTech 2025: Multimodal Hate Speech Detection in Malayalam Using Attention-Driven BiLSTM, Malayalam-Topic-BERT, and Fine-Tuned Wav2Vec 2.0*
Durga Prasad Manukonda, Rohith Gowtham Kodali and Daniel Iglesias

16:00 - 17:30 *byteSizedLLM@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Dravidian Languages Using XLM-RoBERTa and Attention-BiLSTM*
Rohith Gowtham Kodali, Durga Prasad Manukonda and Maharajan Pannakkaran

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *byteSizedLLM@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media Using XLM-RoBERTa and Attention-BiLSTM*
Rohith Gowtham Kodali, Durga Prasad Manukonda and Maharajan Pannakkaran

16:00 - 17:30 *byteSizedLLM@DravidianLangTech 2025: Multimodal Misogyny Meme Detection in Low-Resource Dravidian Languages Using Transliteration-Aware XLM-RoBERTa, ResNet-50, and Attention-BiLSTM*
Durga Prasad Manukonda and Rohith Gowtham Kodali

16:00 - 17:30 *byteSizedLLM@DravidianLangTech 2025: Sentiment Analysis in Tamil Using Transliteration-Aware XLM-RoBERTa and Attention-BiLSTM*
Durga Prasad Manukonda and Rohith Gowtham Kodali

16:00 - 17:30 *SSNCSE@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages*
Sreeja K and Bharathi B

16:00 - 17:30 *YenCS@DravidianLangTech 2025: Integrating Hybrid Architectures for Fake News Detection in Low-Resource Dravidian Languages*
Anusha M D Gowda and Parameshwar R Hegde

16:00 - 17:30 *Girma@DravidianLangTech 2025: Detecting AI Generated Product Reviews*
Girma Yohannis Bade, Muhammad Tayyab Zamir, Olga Kolesnikova, José Luis Oropeza, Grigori Sidorov and Alexander Gelbukh

16:00 - 17:30 *Beyond_Tech@DravidianLangTech 2025: Political Multiclass Sentiment Analysis using Machine Learning and Neural Network*
Kogilavani Shanmugavadivel, Malliga Subramanian, Sanjai R, Mohammed Sameer and Motheeswaran K

16:00 - 17:30 *HTMS@DravidianLangTech 2025: Fusing TF-IDF and BERT with Dimensionality Reduction for Abusive Language Detection in Tamil and Malayalam*
Bachu Naga Sri Harini, Kankipati Venkata Meghana, Kondakindi Supriya, Tara Samiksha and Premjith B

16:00 - 17:30 *Team_Catalysts@DravidianLangTech 2025: Leveraging Political Sentiment Analysis using Machine Learning Techniques for Classifying Tamil Tweets*
Kogilavani Shanmugavadivel, Malliga Subramanian, Subhadevi K, Sowbharanika Janani Sivakumar and Rahul K

16:00 - 17:30 *InnovationEngineers@DravidianLangTech 2025: Enhanced CNN Models for Detecting Misogyny in Tamil Memes Using Image and Text Classification*
Kogilavani Shanmugavadivel, Malliga Subramanian, Poojasree M, Palanimurugan Palanimurugan and Roshini Priya

16:00 - 17:30 *MysticCIOL@DravidianLangTech 2025: A Hybrid Framework for Sentiment Analysis in Tamil and Tulu Using Fine-Tuned SBERT Embeddings and Custom MLP Architectures*
Minhaz Chowdhury, Arnab Laskar, Taj Ahmad and Azmine Toushik Wasi

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *KEC_AI_DATA_DRIFTERS@DravidianLangTech 2025: Fake News Detection in Dravidian Languages*
Kogilavani Shanmugavadivel, Malliga Subramanian, Vishali K S, Priyanka B and Naveen Kumar K

16:00 - 17:30 *KECEmpower@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media*
Malliga Subramanian, Kogilavani Shanmugavadivel, Indhuja V S, Kowshik P and Jayasurya S

16:00 - 17:30 *KEC_AI_GRYFFINDOR@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages*
Kogilavani Shanmugavadivel, Malliga Subramanian, ShahidKhan S, Shri Sashmitha.s and Yashica S

16:00 - 17:30 *KECLinguAIsts@DravidianLangTech 2025: Detecting AI-generated Product Reviews in Dravidian Languages*
Malliga Subramanian, Rojitha R, Mithun Chakravarthy Y, Renusri R V and Kogilavani Shanmugavadivel

16:00 - 17:30 *Dll5143@DravidianLangTech 2025: Majority Voting-Based Framework for Misogyny Meme Detection in Tamil and Malayalam*
Sarbajeet Pattanaik, Ashok Yadav and Vrijendra Singh

16:00 - 17:30 *KEC_AI_VSS_run2@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media*
Kogilavani Shanmugavadivel, Malliga Subramanian, Sathiyaseelan S, Suresh Babu K and Vasikaran S

16:00 - 17:30 *The_Deathly_Hallows@DravidianLangTech 2025: AI Content Detection in Dravidian Languages*
Kogilavani Shanmugavadivel, Malliga Subramanian, Vasantharan K, Prethish G A and Vijayakumaran S

16:00 - 17:30 *SSN_MMHS@DravidianLangTech 2025: A Dual Transformer Approach for Multimodal Hate Speech Detection in Dravidian Languages*
Jahnavi Murali and Rajalakshmi Sivanaiah

16:00 - 17:30 *InnovateX@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Dravidian Languages*
Moogambigai A, Pandiarajan D and Bharathi B

16:00 - 17:30 *KSK@DravidianLangTech 2025: Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments Using Incremental Learning*
Kalaivani K S, Sanjay R, Thissyakkanna S M and Nirenjhanram S K

16:00 - 17:30 *BlueRay@DravidianLangTech-2025: Fake News Detection in Dravidian Languages*
Kogilavani Shanmugavadivel, Malliga Subramanian, Aiswarya M, Aruna T and Jeevaananth S

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *KEC_AI_ZEROWATTS@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages*
Kogilavani Shanmugavadivel, Malliga Subramanian, Naveenram C E, Vishal RS and Srinesh S

16:00 - 17:30 *MNLP@DravidianLangTech 2025: A Deep Multimodal Neural Network for Hate Speech Detection in Dravidian Languages*
Shraddha Chauhan and Abhinav Kumar

16:00 - 17:30 *MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL Based Transformer Models for Detection of Abusive Tamil and Malayalam Text Targeting Women on Social Media*
Md Mizanur Rahman, Srijita Dhar, Md Mehedi Hasan and Hasan Murad

16:00 - 17:30 *MNLP@DravidianLangTech 2025: Transformer-based Multimodal Framework for Misogyny Meme Detection*
Shraddha Chauhan and Abhinav Kumar

16:00 - 17:30 *Code_Conquerors@DravidianLangTech 2025: Deep Learning Approach for Sentiment Analysis in Tamil and Tulu*
Harish Vijay V, Ippatapu Venkata Srichandra, Pathange Omkareshwara Rao and Premjith B

16:00 - 17:30 *KEC_TECH_TITANS@DravidianLangTech 2025: Abusive Text Detection in Tamil and Malayalam Social Media Comments Using Machine Learning*
Malliga Subramanian, Kogilavani Shanmugavadivel, Deepiga P, Dharshini S, Ananthakumar S and Praveenkumar C

16:00 - 17:30 *JustATalentedTeam@DravidianLangTech 2025: A Study of ML and DL approaches for Sentiment Analysis in Code-Mixed Tamil and Tulu Texts*
Ponsubash Raj R, Paruvatha Priya B and Bharathi B

16:00 - 17:30 *KEC_TECH_TITANS@DravidianLangTech 2025:Sentiment Analysis for Low-Resource Languages: Insights from Tamil and Tulu using Deep Learning and Machine Learning Models*
Malliga Subramanian, Kogilavani Shanmugavadivel, Dharshini S, Deepiga P, Praveenkumar C and Ananthakumar S

16:00 - 17:30 *Code_Conquerors@DravidianLangTech 2025: Multimodal Misogyny Detection in Dravidian Languages Using Vision Transformer and BERT*
Pathange Omkareshwara Rao, Harish Vijay V, Ippatapu Venkata Srichandra, Neethu Mohan and Sachin Kumar S

16:00 - 17:30 *YenLP_CS@DravidianLangTech 2025: Sentiment Analysis on Code-Mixed Tamil-Tulu Data Using Machine Learning and Deep Learning Models*
Raksha Adyanthaya and Rathnakara Shetty P

16:00 - 17:30 *LinguAIsts@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media*
Dhanyashree G, Kalpana K, Lekhashree A, Arivuchudar K, Arthi R, Bommineni Sahitya, Pavithra J and Sandra Johnson

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *KEC-Elite-Analysts@DravidianLangTech 2025: Deciphering Emotions in Tamil-English and Code-Mixed Social Media Tweets*
Malliga Subramanian, Aruna A, Anbarasan T, Amudhavan M, Jahaganapathi S and Kogilavani Shanmugavadivel

16:00 - 17:30 *Cyber Protectors@DravidianLangTech 2025: Abusive Tamil and Malayalam Text Targeting Women on Social Media using FastText*
Rohit VP, Madhav M, Ippatapu Venkata Srichandra, Neethu Mohan and Sachin Kumar S

16:00 - 17:30 *LinguAIsts@DravidianLangTech 2025: Misogyny Meme Detection using multi-model Approach*
Arthi R, Pavithra J, Dr G Manikandan, Lekhashree A, Dhanyashree G, Bommineni Sahitya, Arivuchudar K and Kalpana K

16:00 - 17:30 *CUET_Agile@DravidianLangTech 2025: Fine-tuning Transformers for Detecting Abusive Text Targeting Women from Tamil and Malayalam Texts*
Tareque Md Hanif and Md Rashadur Rahman

16:00 - 17:30 *Necto@DravidianLangTech 2025: Fine-tuning Multilingual MiniLM for Text Classification in Dravidian Languages*
Livin Nector Dhasan

16:00 - 17:30 *CUET-823@DravidianLangTech 2025: Shared Task on Multimodal Misogyny Meme Detection in Tamil Language*
Arpita Mallik, Ratnajit Dhar, Udoy Das, Momtazul Arefin Labib, Samia Rahman and Hasan Murad

16:00 - 17:30 *Hermes@DravidianLangTech 2025: Sentiment Analysis of Dravidian Languages using XLM-RoBERTa*
Emmanuel George P, Ashiq Firoz, Madhav Murali, Siranjeevi Rajamanickam and Balasubramanian Palani

16:00 - 17:30 *SSNTrio@DravidianLangTech 2025: Identification of AI Generated Content in Dravidian Languages using Transformers*
J Bhuvana, Mirnalinee T T, Rohan R, Diya Seshan and Avaneesh Koushik

16:00 - 17:30 *SSNTrio@DravidianLangTech 2025: Sentiment Analysis in Dravidian Languages using Multilingual BERT*
J Bhuvana, Mirnalinee T T, Diya Seshan, Rohan R and Avaneesh Koushik

16:00 - 17:30 *NLP_goats@DravidianLangTech 2025: Detecting Fake News in Dravidian Languages: A Text Classification Approach*
Srihari V K, Vijay Karthick Vaidyanathan and Thenmozhi Durairaj

16:00 - 17:30 *NLP_goats@DravidianLangTech 2025: Towards Safer Social Media: Detecting Abusive Language Directed at Women in Dravidian Languages*
Vijay Karthick Vaidyanathan, Srihari V K and Thenmozhi Durairaj

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *HerWILL@DravidianLangTech 2025: Ensemble Approach for Misogyny Detection in Memes Using Pre-trained Text and Vision Transformers*
Neelima Monjusha Preeti, Trina Chakraborty, Noor Mairukh Khan Arnob, Saiyara Mahmud and Azmine Toushik Wasi

16:00 - 17:30 *Cognitext@DravidianLangTech2025: Fake News Classification in Malayalam Using mBERT and LSTM*
Shriya Alladi and Bharathi B

16:00 - 17:30 *NLP_goats_DravidianLangTech_2025__Detecting_AI_Written_Reviews_for_Consumer_Trust*
Srihari V K, Vijay Karthick Vaidyanathan, Mugilkrishna D U and Thenmozhi Durairaj

16:00 - 17:30 *RATHAN@DravidianLangTech 2025: Annaparavai - Separate the Authentic Human Reviews from AI-generated one*
Jubeerathan Thevakumar and Luheerathan Thevakumar

16:00 - 17:30 *DLRG@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages*
Ratnavel Rajalakshmi, Ramesh Kannan, Meetesh Saini and Bitan Mallik

16:00 - 17:30 *Team ML_Forge@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages*
Adnan Faisal, Shiti Chowdhury, Sajib Bhattacharjee, Udoy Das, Samia Rahman, Momtazul Arefin Labib and Hasan Murad

16:00 - 17:30 *codecrackers@DravidianLangTech 2025: Sentiment Classification in Tamil and Tulu Code-Mixed Social Media Text Using Machine Learning*
Lalith Kishore V P, Dr G Manikandan, Mohan Raj M A, Keerthi Vasan A and Aravindh M

16:00 - 17:30 *CUET_Ignite@DravidianLangTech 2025: Detection of Abusive Comments in Tamil Text Using Transformer Models*
MD.Mahadi Rahman, Mohammad Minhaj Uddin and Mohammad Shamsul Arefin

16:00 - 17:30 *CUET_Absolute_Zero@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Malayalam and Tamil Language Using Transformer Models*
Anindo Barua, Sidratul Muntaha, Momtazul Arefin Labib, Samia Rahman, Udoy Das and Hasan Murad

16:00 - 17:30 *MNLP@DravidianLangTech 2025: Transformers vs. Traditional Machine Learning: Analyzing Sentiment in Tamil Social Media Posts*
Abhay Vishwakarma and Abhinav Kumar

16:00 - 17:30 *shimig@DravidianLangTech2025: Stratification of Abusive content on Women in Social Media*
Gersome Shimi, Jerin Mahibha C and Thenmozhi Durairaj

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *SSNTrio@DravidianLangTech2025: LLM Based Techniques for Detection of Abusive Text Targeting Women*
Mirnalinee T T, J Bhuvana, Avaneesh Koushik, Diya Seshan and Rohan R

16:00 - 17:30 *CUET-NLP_MP@DravidianLangTech 2025: A Transformer and LLM-Based Ensemble Approach for Fake News Detection in Dravidian*
Md Minhazul Kabir, Md. Mohiuddin, Kawsar Ahmed and Mohammed Moshiul Hoque

16:00 - 17:30 *CUET-NLP_Big_O@DravidianLangTech 2025: A Multimodal Fusion-based Approach for Identifying Misogyny Memes*
Md. Refaj Hossan, Nazmus Sakib, Md. Alam Miah, Jawad Hossain and Mohammed Moshiul Hoque

16:00 - 17:30 *LexiLogic@DravidianLangTech 2025: Detecting Misogynistic Memes and Abusive Tamil and Malayalam Text Targeting Women on Social Media*
Niranjan Kumar M, Pranav Gupta, Billodal Roy and Souvik Bhattacharyya

16:00 - 17:30 *CUET-NLP_Big_O@DravidianLangTech 2025: A BERT-based Approach to Detect Fake News from Malayalam Social Media Texts*
Nazmus Sakib, Md. Refaj Hossan, Alamgir Hossain, Jawad Hossain and Mohammed Moshiul Hoque

16:00 - 17:30 *LexiLogic@DravidianLangTech 2025: Detecting Fake News in Malayalam and AI-Generated Product Reviews in Tamil and Malayalam*
Souvik Bhattacharyya, Pranav Gupta, Niranjan Kumar M and Billodal Roy

16:00 - 17:30 *SSNTrio @ DravidianLangTech 2025: Hybrid Approach for Hate Speech Detection in Dravidian Languages with Text and Audio Modalities*
J Bhuvana, Mirnalinee T T, Rohan R, Diya Seshan and Avaneesh Koushik

16:00 - 17:30 *Fired_from_NLP@DravidianLangTech 2025: A Multimodal Approach for Detecting Misogynistic Content in Tamil and Malayalam Memes*
Md. Sajid Alam Chowdhury, Mostak Mahmud Chowdhury, Anik Mahmud Shanto, Jidan Al Abrar and Hasan Murad

16:00 - 17:30 *One_by_zero@DravidianLangTech 2025: Fake News Detection in Malayalam Language Leveraging Transformer-based Approach*
Dola Chakraborty, Shamima Afroz, Jawad Hossain and Mohammed Moshiul Hoque

16:00 - 17:30 *CUET_Novice@DravidianLangTech 2025: A Multimodal Transformer-Based Approach for Detecting Misogynistic Memes in Malayalam Language*
Khadiza Sultana Sayma, Farjana Alam Tofa, Md Osama and Ashim Dey

16:00 - 17:30 *teamiic@DravidianLangTech2025-NAACL 2025: Transformer-Based Multimodal Feature Fusion for Misogynistic Meme Detection in Low-Resource Dravidian Language*
Harshita Sharma, Simran Simran, Vajratiya Vajrobol and Nitisha Aggarwal

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *CUET_Novice@DravidianLangTech 2025: Abusive Comment Detection in Malayalam Text Targeting Women on Social Media Using Transformer-Based Models*
Farjana Alam Tofa, Khadiza Sultana Sayma, Md Osama and Ashim Dey

16:00 - 17:30 *SemanticCuetSync@DravidianLangTech 2025: Multimodal Fusion for Hate Speech Detection - A Transformer Based Approach with Cross-Modal Attention*
Md. Sajjad Hossain, Symom Hossain Shohan, Ashraful Islam Paran, Jawad Hossain and Mohammed Moshiul Hoque

16:00 - 17:30 *CUET_Novice@DravidianLangTech 2025: A Bi-GRU Approach for Multiclass Political Sentiment Analysis of Tamil Twitter (X) Comments*
Arupa Barua, Md Osama and Ashim Dey

16:00 - 17:30 *CIC-NLP@DravidianLangTech 2025: Detecting AI-generated Product Reviews in Dravidian Languages*
Tewodros Achamaleh, Tolulope Olalekan Abiola, Lemlem Eyob Kawo, Mikiyas Mebraihtu and Grigori Sidorov

16:00 - 17:30 *One_by_zero@DravidianLangTech 2025: A Multimodal Approach for Misogyny Meme Detection in Malayalam Leveraging Visual and Textual Features*
Dola Chakraborty, Shamima Afroz, Jawad Hossain and Mohammed Moshiul Hoque

16:00 - 17:30 *CUET-NLP_MP@DravidianLangTech 2025: A Transformer-Based Approach for Bridging Text and Vision in Misogyny Meme Detection in Dravidian Languages*
Md. Mohiuddin, Md Minhazul Kabir, Kawsar Ahmed and Mohammed Moshiul Hoque

16:00 - 17:30 *CUET_NetworkSociety@DravidianLangTech 2025: A Transformer-Based Approach to Detecting AI-Generated Product Reviews in Low-Resource Dravidian Languages*
Sabik Aftahee, Tofayel Ahmmed Babu, MD Musa Kalimullah Ratul, Jawad Hossain and Mohammed Moshiul Hoque

16:00 - 17:30 *CUET_NetworkSociety@DravidianLangTech 2025: A Multimodal Framework to Detect Misogyny Meme in Dravidian Languages*
MD Musa Kalimullah Ratul, Sabik Aftahee, Tofayel Ahmmed Babu, Jawad Hossain and Mohammed Moshiul Hoque

16:00 - 17:30 *CUET_NetworkSociety@DravidianLangTech 2025: A Transformer-Driven Approach to Political Sentiment Analysis of Tamil X (Twitter) Comments*
Tofayel Ahmmed Babu, MD Musa Kalimullah Ratul, Sabik Aftahee, Jawad Hossain and Mohammed Moshiul Hoque

16:00 - 17:30 *cantnlp@DravidianLangTech-2025: A Bag-of-Sounds Approach to Multimodal Hate Speech Detection*
Sidney Wong and Andrew Li

16:00 - 17:30 *LexiLogic@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages*
Billodal Roy, Pranav Gupta, Souvik Bhattacharyya and Niranjan Kumar M

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *LexiLogic@DravidianLangTech 2025: Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments and Sentiment Analysis in Tamil and Tulu*
Billodal Roy, Souvik Bhattacharyya, Pranav Gupta and Niranjan Kumar M

16:00 - 17:30 *DLTCNITPY@DravidianLangTech 2025 Abusive Code-mixed Text Detection System Targeting Women for Tamil and Malayalam Languages using Deep Learning Technique*
Habiba A and DR G Aghila

16:00 - 17:30 *Hydrangea@DravidianLanTech2025: Abusive language Identification from Tamil and Malayalam Text using Transformer Models*
Shanmitha Thirumoorthy, Thenmozhi Durairaj and Ratnavel Rajalakshmi

16:00 - 17:30 *CUET_NLP_FiniteInfinity@DravidianLangTech 2025: Exploring Large Language Models for AI-Generated Product Review Classification in Malayalam*
Md. Zahid Hasan, Safiul Alam Sarker, MD Musa Kalimullah Ratul, Kawsar Ahmed and Mohammed Moshiul Hoque

16:00 - 17:30 *NAYEL@DravidianLangTech-2025: Character N-gram and Machine Learning Coordination for Fake News Detection in Dravidian Languages*
Hamada Nayel, Mohammed Aldawsari and Hosahalli Lakshmaiah Shashirekha

16:00 - 17:30 *AnalysisArchitects@DravidianLangTech 2025: BERT Based Approach For Detecting AI Generated Product Reviews In Dravidian Languages*
Abirami Jayaraman, Aruna Devi Shanmugam, Dharunika Sasikumar and Bharathi B

16:00 - 17:30 *AnalysisArchitects@DravidianLangTech 2025: Machine Learning Approach to Political Multiclass Sentiment Analysis of Tamil*
Abirami Jayaraman, Aruna Devi Shanmugam, Dharunika Sasikumar and Bharathi B

16:00 - 17:30 *TEAM_STRIKERS@DravidianLangTech2025: Misogyny Meme Detection in Tamil Using Multimodal Deep Learning*
Kogilavani Shanmugavadivel, Malliga Subramanian, Mohamed Arsath H, Ramya K and Ragav R

16:00 - 17:30 *KCRL@DravidianLangTech 2025: Multi-Pooling Feature Fusion with XLM-RoBERTa for Malayalam Fake News Detection and Classification*
Fariha Haq, Md. Tanvir Ahammed Shawon, Md Ayon Mia, Golam Sarwar Md. Mursalin and Muhammad Ibrahim Khan

16:00 - 17:30 *KCRL@DravidianLangTech 2025: Multi-View Feature Fusion with XLM-R for Tamil Political Sentiment Analysis*
Md Ayon Mia, Fariha Haq, Md. Tanvir Ahammed Shawon, Golam Sarwar Md. Mursalin and Muhammad Ibrahim Khan

16:00 - 17:30 *TensorTalk@DravidianLangTech 2025: Sentiment Analysis in Tamil and Tulu using Logistic Regression and SVM*
K Anishka and Anne Jacika J

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *TeamVision@DravidianLangTech 2025: Detecting AI generated product reviews in Dravidian Languages*
Shankari S R, Sarumathi P and Bharathi B

16:00 - 17:30 *CIC-NLP@DravidianLangTech 2025: Fake News Detection in Dravidian Languages*
Tewodros Achamaleh, Nida Hafeez, Mikiyas Mebraihtu, Fatima Uroosa and Grigori Sidorov

16:00 - 17:30 *CoreFour_IIITK@DravidianLangTech 2025: Abusive Content Detection Against Women Using Machine Learning And Deep Learning Models*
Varun Balaji S, Bojja Revanth Reddy, Vyshnavi Reddy Battula, Suraj Nagunuri and Balasubramanian Palani

16:00 - 17:30 *The_Deathly_Hallows@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages*
Kogilavani Shanmugavadivel, Malliga Subramanian, Vasantharan K, Prethish G A and Santhosh S

16:00 - 17:30 *SSN_IT_NLP@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media*
Maria Nancy C, Radha N and Swathika R

16:00 - 17:30 *LinguAIsts@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media*
Dhanyashree G, Kalpana K, Lekhashree A, Arivuchudar K, Arthi R, Bommineni Sahitya, Pavithra J and Sandra Johnson

16:00 - 17:30 *Celestia@DravidianLangTech 2025: Malayalam-BERT and m-BERT based transformer models for Fake News Detection in Dravidian Languages*
Syeda Alisha Noor, Sadia Anjum, Syed Ahmad Reza and Md Rashadur Rahman

16:00 - 17:30 *Trio Innovators @ DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages*
Radha N, Swathika R, Farha Afreen I, Annu G and Apoorva A

16:00 - 17:30 *Wictory@DravidianLangTech 2025: Political Sentiment Analysis of Tamil X(Twitter) Comments using LaBSE and SVM*
Nithish Ariyha K, Eshwanth Karti T R, Yeshwanth Balaji A P, Vikash J and Sachin Kumar S

16:00 - 17:30 *ANSR@DravidianLangTech 2025: Detection of Abusive Tamil and Malayalam Text Targeting Women on Social Media using RoBERTa and XGBoost*
Nishanth S, Shruthi Rengarajan, S Ananthasivan, Burugu Rahul and Sachin Kumar S

16:00 - 17:30 *Synapse@DravidianLangTech 2025: Multiclass Political Sentiment Analysis in Tamil X (Twitter) Comments: Leveraging Feature Fusion of IndicBERTv2 and Lexical Representations*
Suriya KP, Durai Singh K, Vishal A S, Kishor S and Sachin Kumar S

Saturday, May 3, 2025 (continued)

16:00 - 17:30 *cuetRaptors@DravidianLangTech 2025: Transformer-Based Approaches for Detecting Abusive Tamil Text Targeting Women on Social Media*
Md. Mubasshir Naib, Md. Saikat Hossain Shohag, Alamgir Hossain, Jawad Hossain and Mohammed Moshiul Hoque

16:00 - 17:30 *KEC_AI_BRIGHTRED@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages*
Kogilavani Shanmugavadivel, Malliga Subramanian, Nishdharani P, Santhiya E and Yaswanth Raj E

17:30 - 17:45 ***Meeting, Awards, Closing Remarks***

F^2 (FutureFiction): Detection of Fake News on Futuristic Technology

MSVPJ Sathvik
Raickers AI
Hyderabad
Telangana, India
msvpjsathvik@gmail.com

Venkatesh Velugubantla
Meridian Cooperative
Atlanta
Georgia, USA
venki.v@gmail.com

Ravi Teja Potla
Slalom
Houston
Texas, USA
raviteja.potla@gmail.com

Abstract

There is widespread of misinformation on futuristic technology and society. To accurately detect such news, the algorithms require up-to-date knowledge. The Large Language Models excel in the NLP but cannot retrieve the ongoing events or innovations. For example, GPT and it's variants are restricted till the knowledge of 2021. We introduce a new methodology for the identification of fake news pertaining to futuristic technology and society. Leveraging the power of Google Knowledge, we enhance the capabilities of the GPT-3.5 language model, thereby elevating its performance in the detection of misinformation. The proposed framework exhibits superior efficacy compared to established baselines with the accuracy of 81.04%. Moreover, we propose a novel dataset consisting of fake news in three languages English, Telugu and Tenglish of around 21000 from various sources.

1 Introduction

In the rapidly evolving landscape of futuristic technology, misinformation has become a pervasive and concerning issue. As groundbreaking innovations such as artificial intelligence, quantum computing, and advanced robotics continue to shape the future, the spread of inaccurate or exaggerated information about these technologies can have profound effects(Marche et al., 2023). Misinformation can distort public perceptions, creating unwarranted fears or unrealistic expectations about the capabilities and implications of these technologies. This, in turn, may lead to misguided policy decisions, hinder the adoption of beneficial technologies, or even fuel unnecessary public concerns that impede the responsible development of emerging innovations(Raponi et al., 2022; Wang et al., 2023).

Moreover, misinformation in the realm of futuristic technology can contribute to a lack of trust in scientific advancements and technological progress. When individuals are exposed to sensationalized or

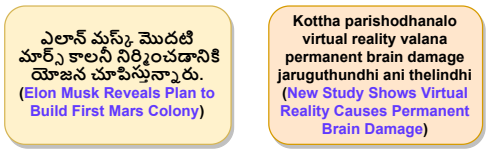


Figure 1: Examples of the fake news related to futuristic technology and society in Telugu and Tenglish. The text in blue color is the translation of the news in English.

inaccurate portrayals of futuristic technologies, it erodes the foundation of public confidence in the scientific community and the technology industry. This erosion of trust can impede collaboration between researchers, policymakers, and the public, hindering the collective efforts needed to navigate the ethical, social, and economic challenges associated with emerging technologies. To address this issue, it is crucial for scientists, technologists, and communicators to prioritize accurate and accessible information, fostering a more informed and discerning public that can engage with the future of technology in a responsible and constructive manner.

To effectively identify and counteract such disinformation, detection algorithms need real-time data to understand the context and verify the accuracy of the information being circulated. Timely updates ensure that the models can recognize and respond to emerging trends, preventing the amplification of false narratives that may contribute to confusion, panic, or even influence public opinion and policy decisions during critical moments in a conflict. To facilitate this up-to-time knowledge update, ontologies and graphs can be employed, offering a structured representation of information that aids in discerning patterns and relationships(Xue and Liu, 2023; Xie et al., 2023).

However, the construction of such ontologies and graphs is a meticulous process that demands time and expertise. Additionally, the rigidity of these structures makes them less adaptable when

transitioning to different domains or subjects. The intricate task of preparing and maintaining these knowledge structures poses a challenge to swiftly respond to evolving scenarios or to seamlessly shift focus to other areas of concern. To overcome these challenges, a pragmatic approach involves leveraging Google's extensive and constantly updated knowledge base. Google serves as a reservoir of real-time information on a myriad of topics, including geopolitical events and war-related developments. By tapping into this vast repository, we can circumvent the time-consuming process of manual ontology creation and instead harness the immediacy and breadth of Google's knowledge.

By integrating Google's dynamic knowledge with the natural language processing (NLP) capabilities of GPT (Generative Pre-trained Transformer), we create a potent synergy. GPT's proficiency in understanding and generating human-like text, coupled with the real-time insights provided by Google, empowers the system to make more informed and timely predictions regarding the authenticity or falsity of information related to war.

This fusion of GPT's linguistic prowess with Google's up-to-date knowledge not only enhances the accuracy of fake news detection but also ensures adaptability to the ever-evolving landscape of information. As a result, this approach not only improves the predictive capabilities in the context of war-related news but also establishes a robust framework that can be extended to different domains, demonstrating a versatility that is crucial in the fast-paced and diverse world of information analysis.

How can we seamlessly integrate Google's knowledge into GPT? One approach involves leveraging Langchain, or alternatively, employing prompting techniques. However, it's crucial to note that these techniques are essentially prompts and might not outperform, especially since GPT isn't explicitly trained to detect fake news. Addressing this necessitates additional training within the context. In this paper, we present a method on how to effectively infuse GPT with knowledge derived from Google, enhancing its capabilities.

The key contributions of our work is as follows:

1. **Novel dataset:** We present a novel dataset with gold human labelled dataset in three languages, Telugu, English and Tenglish.

2. We have implemented baselines on latest approaches like Langchain, GPT-3.5, etc.

3. **Novel Approach:** We present a new algorithm by leveraging Google's knowledge and GPT's capabilities.

2 Related Work

Several approaches have been proposed for detecting and mitigating the spread of fake news across diverse linguistic and thematic domains. Schütz (2023) introduced a disinformation detection method that leverages knowledge infusion through transfer learning and visualizations. Rehm et al. (2018) presented an infrastructure for handling fake news and online media phenomena, incorporating both automatic and manual web annotations. Zhu et al. (2022) proposed a memory-guided multi-view multi-domain fake news detection framework, emphasizing the importance of multi-modal information. Duong et al. (2023) utilized knowledge graph, Datalog, and KG-BERT for fact-checking Vietnamese information.

Ahmed et al. (2022) focused on automatically generating temporally labeled data using positional lexicon expansion for the purpose of estimating the focus time of news articles. Singhal et al. (2022) established FactDrill, a data repository containing fact-checked social media content, facilitating the study of fake news incidents in India. Thaokar et al. (2022) developed a multi-linguistic fake news detector for Hindi, Marathi, and Telugu, emphasizing the importance of linguistic diversity in detection models.

Raja et al. (2023) proposed a method for fake news detection in Dravidian languages using transfer learning with adaptive fine-tuning, addressing linguistic nuances. Yigezu et al. (2023) explored abusive comment detection in Dravidian languages, employing a deep learning approach. Briskilal et al. (2023) introduced an ensemble method for classifying Telugu idiomatic sentences using deep learning models, contributing to the understanding of local linguistic patterns.

Arya et al. (2022) leveraged question answering to understand context-specific patterns in fact-checked articles in the global South. Ren et al. (2023) proposed fake news classification using tensor decomposition and a graph convolutional network. Xie et al. (2023) introduced a knowledge graph-enhanced heterogeneous graph neural network for fake news detection, emphasizing the importance of structured information. Che et al. (2023) proposed tensor factorization with sparse

and graph regularization for fake news detection on social networks.

Han et al. (2021) discussed the generation of fake documents using probabilistic logic graphs, providing insights into potential adversarial techniques. Ding et al. (2022) introduced Metadetector, a meta-event knowledge transfer approach for fake news detection. Zhu et al. (2021) presented a knowledge-enhanced approach for fact-checking and verification, highlighting the role of knowledge graphs. Clark et al. (2021) integrated transformers and knowledge graphs for Twitter stance detection, demonstrating the effectiveness of combining these two powerful techniques.

Our proposed dataset focuses specifically on fake news related to futuristic technology and society, providing a unique thematic perspective. Moreover, our algorithm incorporates a fusion of Google's knowledge and the GPT-3.5 model, offering a novel and robust approach to fake news detection in this distinctive domain. This combination of thematic focus and advanced model integration contributes to the enrichment and diversification of the existing landscape of fake news detection methodologies.

3 Data

Data is sourced from Twitter posts and news articles, with newspapers such as The Hindu, Eenadu, Deccan Chronicle, Sakshi, Andhrajyothi, Times of India, and The Indian Express contributing to the dataset. To uphold anonymity and adhere to ethical considerations, the information collected from both newspapers and social media posts is paraphrased. For the paraphrasing of English and Telugu data, a freely available paraphrase tool (paraphrase-tool.com), accommodating multiple languages, is employed. Specifically for Tenglish data, annotators are tasked with manual paraphrasing. The collected data pertains to three languages: Telugu, English, and Tenglish, all focusing on futuristic technology and society. All the news articles and posts gathered are till the May 2023.

Data Annotation: Our goal was to acquire manual ground-truth labels indicating the presence of a string evidence to claim the information is fake or real. We distributed the collected data in batches to annotators, ensuring that each data point was assessed by multiple annotators to minimize labeling errors. Additionally, we ensured that the same annotator did not review the same pairs across batches.

Table 1: Statistics of the Dataset

Source	label 0	label 1	Overall
Telugu Newspapers	2792	2864	5656
English Newspapers	2136	2386	4522
Twitter (Telugu)	573	655	1228
Twitter (English)	1258	1372	2630
Twitter (Tenglish)	3538	3850	7388
Total	10297	11127	21424

Subsequently, annotators labeled the data, and finally, we aggregated the labels from all annotators into a single label.

A total of 6 journalists working for the Telugu media and are proficient in English are assigned tasks to complete the annotation, including 4 journalists of experience 3 to 5 years and two senior journalists having the experience of 10+ years. To maintain label quality and reduce subjectivity, a minimum of two annotators needed to agree for a label to be included in the dataset. In cases where the first two annotators did not agree, up to three additional annotators were assigned to annotate.

In the labeling of annotators had to choose from the four labels:

1. "True" - The provided information has significant evidence to claim as true news.

2. "Requires Advice from Senior Journalist" - The information provided requires more expertise to decide whether the information is true or false.

3. "Fake" - The provided information contradicts the fact or the information has significant evidence to claim false.

4. "Indeterminate" - There is insufficient evidence to make a clear decision on whether the information is true or false.

Data labelled as "Indeterminate" by both annotators is excluded. Text labelled as "Requires Advice from Senior Journalist" is presented to two senior journalists, who are asked to categorize the information as true, false, or indeterminate. The senior journalists independently provide labels initially, and in cases of conflicting labels, they engage in discussions to resolve difference.

We have computed inter annotator scores for the annotators Krippendorff's Alpha score as met-

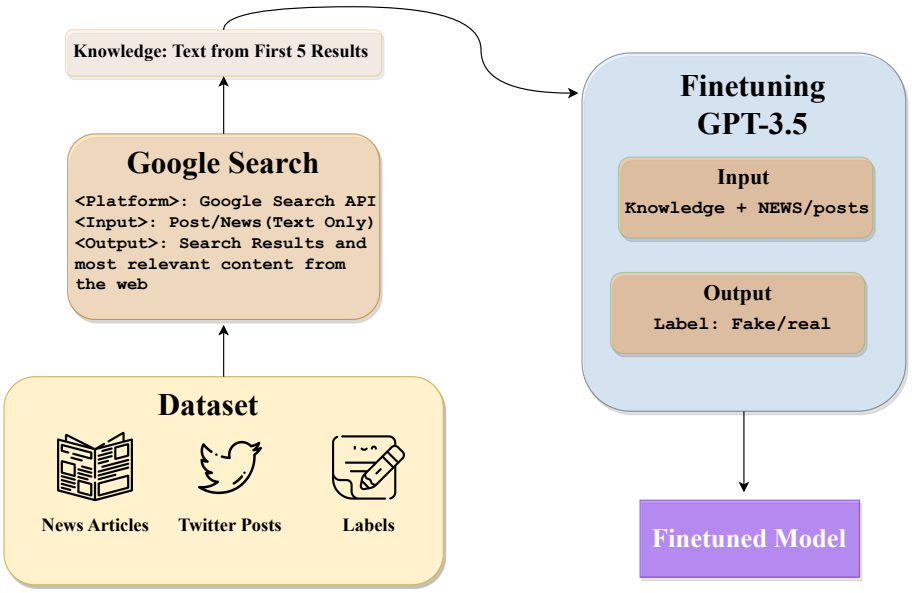


Figure 2: Finetuning of GPT-3.5 by infusing knowledge from Google

ric(Krippendorff, 2011). For the four annotators(a,b,c,d) the scores for each pair of annotators are $\alpha_{ab} = 0.863$, $\alpha_{ac} = 0.837$, $\alpha_{ad} = 0.847$, $\alpha_{bc} = 0.872$, $\alpha_{bd} = 0.861$ and $\alpha_{cd} = 0.854$, . To find out the overall agreement score, the average score for the four annotators , $\alpha = 0.856$. The inter agreement scores for the senior journalists is 0.876. The overall average score, $\alpha_T = 0.866$.

Statistics: Table I illustrates the statistics of the dataset. The dataset analysis highlights the distribution of true and false news across various sources. Telugu and English newspapers contribute a balanced representation, with both categories containing over 2000 instances each. Notably, the Tenglish Twitter category, combining Telugu and English tweets, stands out with a substantial 7388 instances, underscoring its significance as a major source of news content. This Twitter category exhibits a higher volume of both true and false news compared to traditional newspapers.

In total, the dataset comprises 21424 instances, with 10297 instances labeled as true news and 11127 instances labeled as false news. The findings underscore the necessity of source-specific considerations in addressing misinformation, as different platforms exhibit varying levels of reliability. The insights gleaned from this analysis can guide the development of more nuanced and effective strategies for detecting and mitigating misinformation in news content, particularly on dynamic platforms like Twitter.

4 Methodology

4.1 Proposed Algorithm

The proposed algorithm centers around enhancing the capabilities of GPT-3.5 to discern fake news through the integration of information gathered from Google. This strategic approach involves initiating the algorithm by forwarding the input text to Google, retrieving the top five most relevant results. These selected links serve as repositories of crucial information germane to the subject matter of the given news or text. By extracting text from these links, the algorithm gains access to insights encompassing technological advancements and societal developments. This real-time and up-to-date information proves invaluable, particularly terms unfamiliar to GPT-3.5 and tracking developments beyond its training data cut-off in 2021.

The knowledge acquired from these web results becomes an integral part of the fine-tuning process. In this phase, the text obtained from Google is seamlessly integrated with the original news input provided to GPT-3.5, as illustrated in Figure 2. The main goal during the fine-tuning is to enrich the model's understanding by incorporating the wealth of information garnered from Google. This amalgamation enhances the model's grasp of context, allowing it to better comprehend and interpret the intricacies of the information it processes. By training GPT-3.5 with insights from Google, the algorithm seeks to capitalize on the external knowledge to bolster the model's discernment capa-

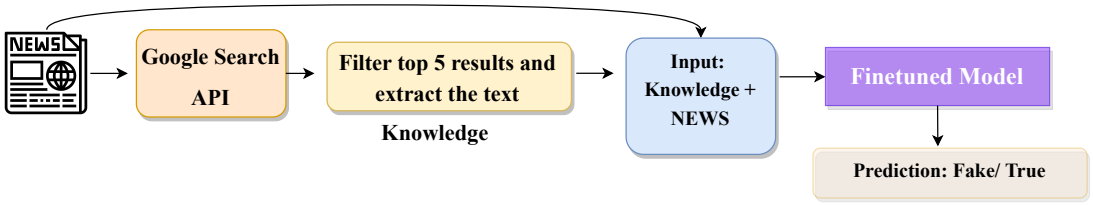


Figure 3: Flowchart depicts the usage of fine-tuned model for testing phase and in real-time applications.

bilities. This additional layer of information equips the model with a broader contextual understanding, enabling it to explore through news content more effectively and identify potential instances of misinformation. In essence, the integration of real-time information from Google serves as a dynamic enhancement strategy, addressing the evolving landscape of information beyond the initial training scope of the model.

Subsequent to the fine-tuning process, the algorithm seamlessly transitions to the analytical phase where the news or post targeted for scrutiny is dispatched to the Google API for information retrieval. This interaction initiates a process where the acquired information is systematically fed into the fine-tuned GPT-3.5 model, as visually depicted in Figure 3. Given that the model has undergone explicit training with the infusion of knowledge from Google, it manifests advanced predictive capabilities in comparison to baseline implementations.

The collaborative synergy between the fine-tuned model and the information retrieved from the Google API underscores a sophisticated approach to tackling the challenges associated with fake news detection. By leveraging external knowledge, the algorithm not only adapts to the evolving landscape of information but also enhances its analytical prowess, contributing to a more robust and effective tool for the detection of fake news and posts in the digital sphere.

To explain in detail we present the mathematical equations. η represents the news from the dataset and N_i the tokens of the news. C is the representation of the label. Assuming G is the notation for the Google search API.

$$News : \eta = \{N_1, N_2, \ldots, N_n\} \tag{1}$$

$$c = \begin{cases} 1, & \text{if } \textit{fake news} \\ 0, & \text{if } \textit{true news} \end{cases} \tag{2}$$

$$\pi = G(N_1, N_2, \ldots, N_n) \tag{3}$$

π represents the results obtained from the Google search API with τ_i as the links. Then the text K_i is extracted from τ_i using ε function while extracts the text from the web links.

$$\pi = \{\tau_1, \tau_2, \ldots, \tau_m\} \tag{4}$$

$$K_i = \varepsilon(\tau_i) \tag{5}$$

$$K_i = \{t_1, t_2, \ldots, t_o\} \tag{6}$$

t_i are the tokens of the text K_i. From the obtained web results the text from first five results is taken and denoted as K_G. The knowledge in addition with the news is fine-tuned on GPT-3.5 with W, B as parameters of the model with loss function L_G and F represents the fine-tuned model.

$$K_G = \{K_1, K_2, \ldots, K_5\} \tag{7}$$

$$F(W,B) = \arg\min_{W,B} L_G(\{K_1+K_2+K_3+K_4+K_5+\eta\}, c) \tag{8}$$

η_t is the news to be predicted with tokens N_{t_i}. The news is input to the Google search and the knowledge extracted is K_{G_t}.

$$\eta_t = \{N_{t_1}, N_{t_2}, \ldots, N_{t_n}\} \tag{9}$$

$$K_{G_t} = \{K_{t_1}, K_{t_2}, \ldots, K_{t_5}\} \tag{10}$$

$$P_f = F(\{K_{t_1} + K_{t_2} + K_{t_3} + K_{t_4} + K_{t_5} + \eta\}) \tag{11}$$

The knowledge and news is input to the fine-tuned model F and the models outputs the prediction P_f.

4.2 Baselines

The data is multilingual, consists of two different languages and a mixed language. So, we have opted multilingual baselines. So, that it could be suitable to evaluate. The implemented baselines

are: (i) GPT 3.5 (Chen et al., 2023); (ii) GPT 3 (Brown et al., 2020);(iii) LLAMA 2(Touvron et al., 2023); (iv) multilingualBERT(Pires et al., 2019); (v)XLM-RoBERTa(Conneau et al., 2020), (vi) Integrating Google and GPT using Langchain(IGL) and (vii) few shot prompting with GPT-3.5.

For the implementation of IGL we have utilised the prompting technique the prompt is "Predict whether the following is fake news or not?: $\backslash n \backslash n$ (The news/post)". For few shot prompting technique, we have prompted the GPT model by providing with five random examples from the dataset.

For the BERT-like models we have used Google Colab free GPU. For LLAMA 2 7B, 13B we have used Nvidia GPU of 108GB RAM. For the GPT models we have utilised the OpenAI API for fine-tuning and few shot prompting of the GPT-3.5.The hyperparameters used for the baselines are epoch 5, learning rate 2e-5, weight decay 0.01, frequency penalty 0, presence penalty 0.

5 Experimental Results

Table 2 presents the experimental results for the experiments performed in this study. TeluguBERT, mBERT, and XLM RoBERTa exhibited competitive performance in the detection of fake news, with accuracy values of 62.37%, 68.15%, and 69.82%, respectively. Among these, XLM RoBERTa achieved the highest accuracy. This might be because RoBERTa is multilingual and as it is enhanced form of BERT. Among the few-shot learning models, Few shot GPT-4 outperformed Few shot GPT-3.5 and IGL, achieving an accuracy of 43.57%. The latter two models demonstrated accuracy values of 41.86% and 51.49%, respectively. The IGL performed better than other prompting techniques this is because of accessing web and gains relevant up to date information.

The LLAMA models, LLAMA 2 7B and LLAMA 2 13B, exhibited superior performance compared to the previous models, achieving accuracy values of 74.15% and 75.86%, respectively. Among the GPT-3 models, GPT 3 Davinci demonstrated the highest accuracy of 75.91%, surpassing GPT 3 Babbage, GPT 3 Curie, and GPT 3 Ada, which achieved accuracy values of 74.57%, 74.36%, and 73.93%, respectively. GPT 3.5 also performed well, with an accuracy of 76.48%. As it is a LLM, pretrained on huge textual data and fine-tuned for detection of the fake news it performed better. The proposed algorithm performed much better than the baselines implemented the main reason is the algorithm learned accessing web, extracting knowledge and detecting the fake news.

Table 2: Test results: Detection of Fake News

Model	Precision	Recall	Accuracy
TeluguBERT	60.27	63.79	62.37
mBERT	65.75	69.56	68.15
XLM RoBERTa	66.72	70.16	69.82
Few shot GPT-3.5	40.62	43.73	41.86
Few shot GPT-4	41.40	43.82	43.57
IGL	50.13	52.32	51.49
LLAMA 2 7B	72.90	76.36	74.15
LLAMA 2 13B	73.61	77.57	75.86
GPT 3 Ada	70.67	74.20	73.93
GPT 3 Babbage	72.45	75.39	74.57
GPT 3 Curie	74.86	73.11	74.36
GPT 3 Davinci	79.26	72.43	75.91
GPT 3.5	74.51	77.06	76.48
Proposed method	79.83	82.17	**81.04**

6 Discussion

In the realm of fake news detection within the futuristic technology landscape, our proposed algorithm, leveraging the fine-tuned GPT-3.5 model with knowledge infusion from Google, outperforms other baselines that also integrate GPT-3.5 but lack dedicated fine-tuning for the specific task of fake news detection. The effectiveness of our approach is evident in its nuanced understanding of language, real-time information retrieval capabilities, and advanced contextual analysis.

One notable strength of our algorithm lies in its ability to discern speculative or sensationalized content that often eludes other baselines. For instance, when faced with a headline proclaiming "Quantum Computing Breakthrough Enables Time Travel," our algorithm excels at cross-referencing the information with recent scientific literature, expert opinions, and official announcements. The fine-tuning process ensures that it recognizes the nuances in language that may signal speculative claims, allowing it to accurately identify potential misinformation where baselines may fall short. Moreover, The fine-tuning process also equips the algorithm with a nuanced understanding of language and context, enhancing its ability to detect subtly misleading information. Consider the headline "AI Singularity Imminent: Experts Warn of Global Catastrophe." Baseline models, integrated with GPT-3.5 but lacking specific fine-tuning, may not grasp the hyperbolic nature of the claim. Our algorithm, having

learned from a multitude of sources, recognizes the speculative tone and lack of substantiated evidence, contributing to a more accurate identification of this headline as potential misinformation.

Additionally, the proposed algorithm demonstrates superior performance in evaluating the credibility of news related to emerging technologies, such as blockchain or artificial intelligence. For instance, when presented with a headline asserting "Blockchain-Powered Flying Cars Set to Hit the Market Next Year", our algorithm can thoroughly analyze the feasibility of such a claim by checking for official statements from industry experts, regulatory approvals, and technological advancements. In contrast, baselines without dedicated fine-tuning for fake news detection may struggle to differentiate between credible and misleading information, relying on general language understanding without the nuanced focus our algorithm provides. Furthermore, in scenarios involving space exploration and extraterrestrial claims, our algorithm's real-time web scraping capabilities ensure that it can access the latest information from reputable sources. For example, when confronted with the headline "NASA Confirms Alien Life on Mars", our algorithm excels at cross-referencing this information with official statements and recent research findings. The dedicated fine-tuning for fake news detection enhances its ability to discern credible sources, enabling it to raise red flags when faced with sensationalized claims, a capability that might be lacking in baselines relying solely on GPT-3.5.

In the cases where knowledge retrieved from Google is incorrect: The Google is not always correct, sometimes we find blogs containing misinformation or fake news. The proposed algorithm performs much better compared to the baselines in this case. The IGL have false positives as they are context-based. As the proposed approach is fine-tuned on data, during the training phase there were data points where the knowledge from the web is incorrect/fake but where the label is true news, during these cases the web results contradicts with label, thereby creating confusion when IGL is used. As the model is fine-tuned it performed well on these cases as well.

Error Analysis:

While our proposed algorithm demonstrates notable strengths in fake news detection within the futuristic technology domain, there are few scenarios it made errors. The algorithm face challenges in distinguishing between legitimate speculation and misinformation in a rapidly evolving field. For example, if a headline speculates on the potential future capabilities of a nascent technology, such as "Experts Predict AI Will Achieve Consciousness Within a Decade," the algorithm struggles to differentiate between speculative but informed predictions and baseless claims. Balancing the understanding of speculative language while avoiding false positives poses a persistent challenge.

Another potential source of error arises when the algorithm encounters news that is related to emerging technologies with limited resources in Telugu. In this scenario the proposed algorithm showed lower performance. There are 6 predictions incorrect for every 10 posts.

From close examination of the predictions we found that the algorithm struggles to detect posts in Tenglish language. This might be because the GPT-3.5 would not have been pretrained on the Tenglish language and therefore feels difficult to understand and detect the fake news. Data augmentation or pertaining on Tenglish language would help in improving the overall performance of the model.

7 Conclusion and Future Work

In conclusion, this study addresses the issue of misinformation in the context of futuristic technology and society. Acknowledging the limitations of existing algorithms, particularly in their inability to incorporate real-time information, we proposed a novel methodology that combines the strengths of Large Language Models, specifically GPT-3.5, with the dynamic knowledge base provided by Google Knowledge. By leveraging this synergy, our framework achieved a commendable accuracy of 81.04% in detecting fake news.

The future work involves scaling the dataset to other language like Hindi, Tamil and other Indic languages. We would also like to develop a pretraining model especially for Tenglish as it is expected to perform better. We would like to develop a dataset in Telugu that also involves the fake news on investments.

Limitations

This approach is specifically designed for handling textual data, ensuring optimized performance for text-based processing. Since it focuses exclusively on text, image data is not included in the dataset,

allowing for a more streamlined and efficient analysis. To leverage advanced AI capabilities, we utilize Google API and OpenAI models, which operate on a structured billing model. This aligns with standard industry practices for accessing state-of-the-art machine learning services. While these models are closed-source, they provide reliable and high-quality performance for text processing.

Ethics Statement

Our primary goal is to detect fake news while ensuring that the reputation of sources remains unaffected. To maintain anonymity, we have rephrased the collected data, preventing any potential reputational impact on sources or users. Additionally, we strongly oppose any misuse of the dataset for generating or spreading fake news.

References

Usman Ahmed, Jerry Chun-Wei Lin*, and Vicente Garcia Diaz. 2022. Automatically temporal labeled data generation using positional lexicon expansion for focus time estimation of news articles. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Arshia Arya, Saloni Dash, Syeda Zainab Akbar, Joyojeet Pal, and Anirban Sen. 2022. Poster: Leveraging question answering to understand context specific patterns in fact checked articles in the global south. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, pages 628–631.

J Briskilal, Ch VM Sai Praneeth, Ch Chaitanya, M Jaya Karthik, and P Purnachandra Reddy. 2023. An ensemble method to classify telugu idiomatic sentences using deep learning models. In *2023 International Conference on Inventive Computation Technologies (ICICT)*, pages 65–71. IEEE.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Hangjun Che, Baicheng Pan, Man-Fai Leung, Yuting Cao, and Zheng Yan. 2023. Tensor factorization with sparse and graph regularization for fake news detection on social networks. *IEEE Transactions on Computational Social Systems*.

Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks. *arXiv preprint arXiv:2303.00293*.

Thomas Clark, Costanza Conforti, Fangyu Liu, Zaiqiao Meng, Ehsan Shareghi, and Nigel Collier. 2021. Integrating transformers and knowledge graphs for twitter stance detection. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 304–312.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yasan Ding, Bin Guo, Yan Liu, Yunji Liang, Haocheng Shen, and Zhiwen Yu. 2022. Metadetector: Meta event knowledge transfer for fake news detection. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(6):1–25.

Huong T Duong, Van H Ho, and Phuc Do. 2023. Fact-checking vietnamese information using knowledge graph, datalog, and kg-bert. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(10):1–23.

Qian Han, Cristian Molinaro, Antonio Picariello, Giancarlo Sperli, Venkatramanan S Subrahmanian, and Yanhai Xiong. 2021. Generating fake documents using probabilistic logic graphs. *IEEE Transactions on Dependable and Secure Computing*, 19(4):2428–2441.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Claudio Marche, Ilaria Cabiddu, Christian Giovanni Castangia, Luigi Serreli, and Michele Nitti. 2023. Implementation of a multi-approach fake news detector and of a trust management model for news sources. *IEEE Transactions on Services Computing*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Simone Raponi, Zeinab Khalifa, Gabriele Oligeri, and Roberto Di Pietro. 2022. Fake news propagation: a review of epidemic models, datasets, and insights. *ACM Transactions on the Web (TWEB)*, 16(3):1–34.

Georg Rehm, Julian Moreno-Schneider, and Peter Bourgonje. 2018. Automatic and manual web annotations in an infrastructure to handle fake news and other online media phenomena. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Qingyun Ren, Bingyin Zhou, Dongli Yan, and Wei Guo. 2023. Fake news classification using tensor decomposition and graph convolutional network. *IEEE Transactions on Computational Social Systems*.

Mina Schütz. 2023. Disinformation detection: Knowledge infusion with transfer learning and visualizations. In *European Conference on Information Retrieval*, pages 468–475. Springer.

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 1322–1331.

Chetana B Thaokar, Mayur Rathod, Shayeek Ahmed, Jitendra Kumar Rout, and Minakhi Rout. 2022. A multi-linguistic fake news detector on hindi, marathi and telugu. In *2022 OITS International Conference on Information Technology (OCIT)*, pages 324–329. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jinxia Wang, Stanislav Makowski, Alan Cieślik, Haibin Lv, and Zhihan Lv. 2023. Fake news in virtual community, virtual society, and metaverse: A survey. *IEEE Transactions on Computational Social Systems*.

Bingbing Xie, Xiaoxiao Ma, Jia Wu, Jian Yang, and Hao Fan. 2023. Knowledge graph enhanced heterogeneous graph neural network for fake news detection. *IEEE Transactions on Consumer Electronics*.

Xingsi Xue and Wenyu Liu. 2023. Integrating heterogeneous ontologies in asian languages through compact genetic algorithm with annealing re-sample inheritance mechanism. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3):1–21.

Mesay Gemeda Yigezu, Selam Kanta, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Habesha@ dravidianlangtech: Abusive comment detection using deep learning approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 244–249.

Biru Zhu, Xingyao Zhang, Ming Gu, and Yangdong Deng. 2021. Knowledge enhanced fact checking and verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3132–3143.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.

TSD: Towards Computational Processing of Tamil Similes – A Tamil Simile Dataset

Aathavan Nithiyananthan, Jathushan Raveendra, Uthayasanker Thayasivam
Department of Computer Science and Engineering, University of Moratuwa,
Sri Lanka
{aathavan.20, jathushan.20, rtuthaya}@cse.mrt.ac.lk

Abstract

A simile is a powerful figure of speech that makes a comparison between two different things via shared properties, often using words like "like" or "as" to create vivid imagery, convey emotions, and enhance understanding. However, computational research on similes is limited in low-resource languages like Tamil due to the lack of simile datasets. This work introduces a manually annotated Tamil Simile Dataset (TSD) comprising around 1.5k simile sentences drawn from various sources. Our data annotation guidelines ensure that all the simile sentences are annotated with the three components, namely *tenor*, *vehicle*, and *context*. We benchmark our dataset for simile interpretation and simile generation tasks using chosen pre-trained language models (PLMs) and present the results. Our findings highlight the challenges of simile tasks in Tamil, suggesting areas for further improvement. We believe that TSD will drive progress in computational simile processing for Tamil and other low-resource languages, further advancing simile related tasks in Natural Language Processing.

1 Introduction

A simile is a figure of speech that explicitly compares two different things by saying that one thing is like another, so it typically contains comparison expressions such as "like" and "as" (Paul, 1970). Similes allow people to create vivid images and convey emotions in ways that literal language cannot. Computational processing of similes is gaining attention in Natural Language Processing (NLP) research which enables the development of more engaging conversational systems (Zheng et al., 2020), creative writing tools (Zhang et al., 2021) and also enhances applications in sentiment analysis (Ge et al., 2023).

Research on simile processing is limited compared to other areas in NLP (Chakrabarty et al., 2022). Early research lacked dedicated datasets which made many researchers to create small, task-specific ones as the field evolved (Ge et al., 2023). The simile identification task has achieved significant advances, but other tasks have yet to gain more traction due to the lack of specifically designed annotated data (Ge et al., 2023). There is still room for the development of theories that abstractly explain the connection between the compared elements in similes (Lai and Nissim, 2024). Due to these constraints, simile related tasks are still challenging for high-resourced languages like English and Chinese in which most of the research in similes is centered.

[இமை]$_{\text{VEHICLE}}$ போலக் [காக்கிறான்]$_{\text{CONTEXT}}$ [கடவுள்]$_{\text{TENOR}}$. [God]$_{\text{TENOR}}$ [protects]$_{\text{CONTEXT}}$ like an [eyelash]$_{\text{VEHICLE}}$.
அவள் [உடம்பு]$_{\text{TENOR}}$ காற்றில் ஆடிய [மரங்களைப்]$_{\text{VEHICLE}}$ போல் [ஆடியது]$_{\text{CONTEXT}}$. Her [body]$_{\text{TENOR}}$ [swayed]$_{\text{CONTEXT}}$ like [trees]$_{\text{VEHICLE}}$ dancing in the wind.
[அலைகள்]$_{\text{VEHICLE}}$ போலவே [மோதும்]$_{\text{CONTEXT}}$ உந்தன் [ஞாபகம்]$_{\text{TENOR}}$. Your [memory]$_{\text{TENOR}}$ [strikes]$_{\text{CONTEXT}}$ like [waves]$_{\text{VEHICLE}}$.

Table 1: Examples of Tamil simile sentences and components.

Considering low-resourced languages, research in computational simile processing is extremely limited. There are no large datasets or established resources, and only a few studies have focused on figurative language like similes in these languages. Tamil is a language with unique cultural and linguistic expressions due to its agglutinative nature and complex morphological structures (Keane, 2004). Languages like Tamil are often overlooked in NLP research due to the lack of sufficient resources like annotated datasets and tools.

The creation of simile datasets has improved over time with different labeling methods as research and tasks developed (Ge et al., 2023). Different simile tasks require datasets with specifically annotated components of similes. Currently, it has become standard to annotate all the components, as this can be used across tasks (Yang et al., 2023; Shao et al., 2024a).

In this work, we present Tamil Simile Datset (TSD), a simile dataset annotated manually with all the components of the simile. The contributions of this paper are:

1. We present the Tamil Simile Dataset (TSD), which is the first simile dataset for the Tamil language.

2. Our monolingual dataset contains 1520 sentences all annotated with TENOR, VEHICLE, and CONTEXT.

3. We evaluate our dataset for Simile interpretation and Simile Generation tasks using chosen pre-trained language models and present the results.

2 Background

A simile (உவமையணி) is a figure of speech in which one concept is described in terms of another known concept that shares similar properties, typically using comparators like "like" (போல) to emphasize the comparison. In a simile, the word or concept which is being described is the TENOR (உவமேயம்). The word or concept used to describe the TENOR is the VEHICLE (உவமானம்). VEHICLE is a component that brings imagery or qualities to mind for comparison. Additionally, CONTEXT (பொதுத்தன்மை) is the property through which the comparison is made. Examples of simile sentences and annotated components are shown in Table 1.

3 Related Works

3.1 Simile Datasets

Several datasets have been developed in English and Chinese, which are high-resourced languages (Joshi et al., 2020) to support research in simile-related tasks. Self Labeled Simile (SLS) dataset (Chakrabarty et al., 2020) and the Writing Polishment Similes (WPS) dataset (Zhang et al., 2021) consists of automatically annotated sentences and were used for simile generation tasks. Chinese Metaphor (CM) dataset (Su et al., 2016), CMC dataset (Li et al., 2022), and MSD dataset (Ma et al., 2023) were annotated with TENOR and VEHICLE. These datasets were used for interpretation and generation tasks. MCP dataset (He et al., 2022), GraCe dataset (Yang et al., 2023), and the most recent CMDAG dataset (Shao et al., 2024a) are annotated with all three components and are also utilized for both simile interpretation and simile generation tasks.

Research on similes remains limited in Dravidian languages (Paul et al., 2024). In low-resource languages, small-scale efforts exist, such as simile generation in Afrikaans (van Heerden and Bas, 2021), and recently, a Malayalam simile dataset for identification has been developed (Paul et al., 2024). Elanchezhiyan et al. (2014) analyzed Tamil song lyrics to identify simile patterns and attributes for automatic simile generation, but no dedicated dataset was released, and further details are unavailable.

3.2 Tasks in Simile Processing

Simile interpretation and simile generation are the two main directions of the simile study (Yu and Wan, 2019).

Simile interpretation task focuses on identifying shared properties between the TENOR and VEHICLE. Early methods relied on word embeddings to measure semantic similarity (Zheng et al., 2020; Bar et al., 2022), but recent work has integrated knowledge bases like ConceptNet (Gero and Chilton, 2019; Stowe et al., 2021). PLMs have further refined interpretation by capturing implicit meanings without predefined rules (Su et al., 2017). Ma et al. (2023) introduced a task where models predict the shared property in a simile, while Chen et al. (2022) used masked language modeling (MLM) to predict missing simile elements.

Simile generation task involves constructing simile expressions. Recent approaches fine-tune pre-trained language models such as GPT-2 (Li et al., 2022) or BART (Lewis, 2019) for this task. Knowledge-driven methods frame it as knowledge graph completion, generating VEHICLEs based on relational context (Song et al., 2020). Chen et al. (2022) refined PLM-based simile matching, while Yang et al. (2023) used CBART (Shao et al., 2024b) with multiple constraints for Chinese simile generation. Ma et al. (2023) extended the task to dialogue systems, requiring models to select

appropriate VEHICLEs. Recent research underscores the importance of PLMs in improving both simile interpretation and generation tasks.

4 Tamil Simile Dataset

In this section, we present the collection, annotation, and statistics of our manually annotated Tamil Simile Dataset.

4.1 Data collection

We collected data from various sources. We used Wikisource API[1] to get random articles from Tamil Wikisource and extracted the article contents which contained Tamil simile comparators such as "போல" (Pola "like") or "போன்ற" (Pondra "like"'). Additionally, we extracted texts which contained morphemes of "போல" such as "போல்" (Pol), "போலே" (Pole) and "போலும்" (Polum)—all of which convey the meaning "like". This included similes from various kinds of literature, such as old Tamil scripts like Kambaramayanam and Tamil poems, stories, and essays. We also extracted similes from Tamil song lyrics, as similes are most frequently used in Tamil songs. We collected songs from tamil2lyrics.com[2] and extracted songs that contained simile comparators as above.

4.2 Data annotation

We employed 10 annotators to extract meaningful sentences from the collected data. In the Tamil language, not all the sentences that contain the comparator "போல" are similes. For example, consider the sentence "கதவு அடைத்து உட்புறமாகத் தாழிட்டிருப்பது போல தெரிந்தது" ("Kathavu adaittu utpuramaga thazhittiruppathu pola therinthathu") (translates to: "The door seemed to be locked and slammed inwards"). Here "போல" is used to convey a state of appearance rather than a direct comparison. So our annotators first extract sentences that are similes and disregard literal sentences.

Sentences annotated as similes are forwarded to the next stage of the annotation. In this stage, another 4 annotators annotated the VEHICLE, TENOR, and the CONTEXT of the sentences. When the annotators extract a sentence as a simile sentence, it will have the comparator and the VEHICLE word by default. So, annotators are asked to annotate the VEHICLE (it can be word/phrase/sentence) first. The next step is to annotate the TENOR and CONTEXT if it is found in the sentence. If not, we instructed the annotators to annotate the TENOR and CONTEXT (both can be a word/phrase/sentence). We asked the annotators to disregard confusing cases where it is difficult to find TENOR or CONTEXT. In this way, we were able to ensure all the simile sentences in our dataset were annotated with all three components. Our annotation process is shown in Figure 1. Before the components annotation phase, a training session was conducted, and annotators were trained with examples and instructed with Tamil simile principles. A set of 50 sample sentences sourced from various genres was given to all 4 annotators for component annotation to check the reliability of their annotation. We computed the inner-annotator agreement of simile component annotation via Krippendorff's alpha (Krippendorff, 2011). The overall agreement rate was found to be 0.78. Statistics of TSD are shown in Table 2.

Measurement	Value
# Simile Sentences	1520
# Distinct Tenors	706
# Distinct Vehicles	1042
# Distinct Contexts	1077
Average # Words per Sentence	6

Table 2: Statistics of the dataset.

5 Tasks

In this section, we introduce 2 tasks for our Tamil simile dataset, including the definition of the tasks, the baselines, evaluation metrics, experimental results, and analysis.

5.1 Simile Interpretation/Generation Tasks

Following prior work on simile interpretation (Song et al., 2020; Zheng et al., 2020; He et al., 2022; Chen et al., 2022; Shuhan et al., 2023) and simile generation (Song et al., 2020; Chen et al., 2022; Shuhan et al., 2023), we define Simile Interpretation/Generation (SI/SG) as a fill-mask objective task. We evaluate the models on 100 samples from our dataset.

For the simile interpretation task, we remove the CONTEXT from the simile sentence and replace it with a blank. The model is required to generate the missing CONTEXT. Similarly, for the simile generation task, we remove the VEHICLE from the

[1] https://ta.wikisource.org/w/api.php
[2] https://www.tamil2lyrics.com/

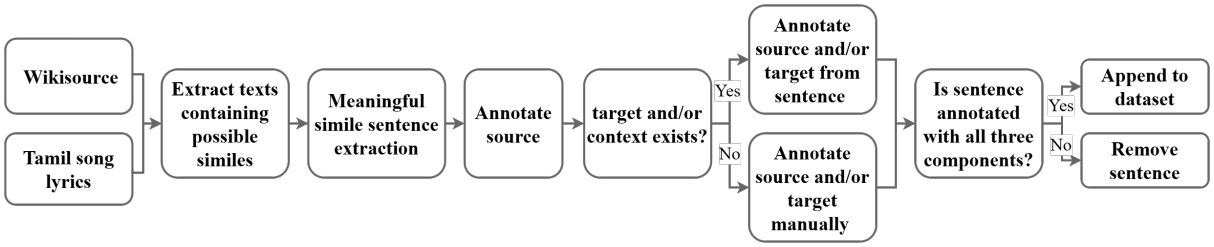


Figure 1: The data annotation process.

Model	Interpretation			Generation		
	MRR↑	Hits@5↑	Hits@10↑	MRR↑	Hits@5↑	Hits@10↑
TamilBERT	0.026	0.05	**0.07**	**0.200**	**0.27**	**0.38**
IndicBERT v2	**0.033**	**0.06**	**0.07**	0.122	0.20	0.30
MuRIL	0.027	**0.06**	0.06	0.140	0.21	0.24
XLM-RoBERTa	0.003	0.01	0.01	0.030	0.05	0.06

Table 3: Simile interpretation and generation results (MRR, Hits@5, Hits@10).

simile sentence, leaving a blank, and the model needs to generate an appropriate VEHICLE. In both cases, we extract the top 10 predicted words from the model.

We fine-tune TamilBERT (Joshi, 2022), IndicBERT v2 (Doddapaneni et al., 2023), MuRIL (Khanuja et al., 2021), and XLM-RoBERTa (Conneau et al., 2019) on the Tamil Simile Dataset. These baselines are chosen due to their strong performance in Dravidian and multilingual NLP tasks, particularly in low-resource settings.

The performance of the models is evaluated using Mean Reciprocal Rank (MRR), Hits@5, and Hits@10. MRR measures the average of the reciprocal ranks of the first correct prediction, providing insight into how well models rank the correct completion. Hits@5 and Hits@10 measure the proportion of cases where the correct word appears within the top 5 and top 10 predictions.

6 Results and Discussion

Table 3 presents the results of simile interpretation and generation tasks. Simile interpretation task yielded lower results compared to the generation task, which may be attributed to the structural characteristics of Tamil simile sentences, where contextual information is sometimes omitted. This aspect requires further investigation. The TamilBERT model achieved relatively high scores in the simile generation task, indicating that a monolingual model trained specifically on Tamil data can be more effective for simile processing in the Tamil language. Additionally, models pre-trained on Indian languages, such as IndicBERT v2 and MuRIL, demonstrated reasonable performance in simile generation. In contrast, XLM-RoBERTa, a multilingual model trained on 100 languages, exhibited weaker performance in simile-related tasks. These findings highlight the impact of language-specific pretraining in low-resource NLP, particularly for complex tasks like simile processing.

Interestingly, the simile interpretation task showed significant improvement during the fine-tuning phase. Initially, the models generated irrelevant tokens such as “##ாதே”, “-”, and “"##ுது”. After fine-tuning, the predictions were contextually appropriate, including words like “பறக்கும்” (flying), “வளைந்த” (curved), and “அழகான” (beautiful). However, we found that many error predictions occurred when the CONTEXT was not a noun or when morphemes complicated interpretation. Further investigation into the effect of tokenization on this task could provide deeper insights into these behaviors. For the generation task, models struggled to predict words that are not so commonly occurring in Tamil language. Exploring alternative fine-tuning techniques may improve the model’s ability to generate more relevant predictions.

Our results and findings indicate that both simile interpretation and generation are challenging for the Tamil language. This can be attributed to Tamil’s linguistic complexities, which make these tasks more difficult compared to languages

with simpler structures. These challenges present valuable opportunities for future research, and the Tamil Simile Dataset (TSD) can serve as a valuable resource for advancing simile processing in low-resource languages.

7 Conclusion

We present a manually annotated Tamil simile dataset (TSD) comprising 1520 simile sentences sourced from a wide range of Tamil literary forms, including poems, short stories, articles, and song lyrics. Our dataset annotators achieved inter-annotator agreement of 0.78, underscoring the reliability of our dataset. We also benchmark our dataset for simile interpretation and simile generation tasks using pre-trained language models. Our results show that simile-related tasks are challenging for Tamil Language. This shows that our dataset has great potential to help improve the understanding and creation of Tamil similes.

8 Limitations

When annotators annotate components TENOR and/or CONTEXT that are not in the original simile sentence manually, there is a possibility of multiple suitable TENORs and/or CONTEXTs for that simile. However, in our dataset, the appropriate one, as determined by the annotators is annotated. The 100 examples we used are sentences that had VEHICLE and CONTEXT within them. Sentences from our dataset which are sourced from tamil2lyrics.com, comprise song lyrics from the 1950s to 2023. This covers a wide range of timelines and songs, though not every song is included. In addition, our dataset consists of different Tamil literary forms such as poems, articles, and other literary sentences extracted from Wikisource. Our dataset is limited in terms of coverage as we could only get sentences from the extracted pages returned by Wikisource API. While Tamil has been rich in figurative language since ancient times, its usage has evolved over time. Expanding simile datasets to include more classical and historical Tamil literature would enhance coverage and further improve computational simile processing in Tamil.

Acknowledgments

This research was funded by the University of Moratuwa Senate Research Committee (SRC) grant SRC/ST/2024/44.

References

Kfir Bar, Nachum Dershowitz, and Lena Dankin. 2022. Metaphor interpretation using word embeddings. *Computación y Sistemas*, 26(3):1301–1311.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. *arXiv preprint arXiv:2009.08942*.

Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Jiashu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yijiang Chen, and Chang Su. 2022. Probing simile knowledge from pre-trained language models. *arXiv preprint arXiv:2204.12807*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

K. Elanchezhiyan, E. Tamil Selvi, N. Revathi, G. P. Shanthi, S. Shireen, and Madhan Karky. 2014. Simile generation. In *Proceedings of the 13th International Tamil Internet Conference*.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(Suppl 2):1829–1895.

Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? *arXiv preprint arXiv:2203.08452*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Elinor Keane. 2004. Tamil. *Journal of the International Phonetic Association*, 34(1):111–116.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Huiyuan Lai and Malvina Nissim. 2024. A survey on automatic generation of figurative language: From rule-based systems to large language models. *ACM Computing Surveys*, 56(10):1–34.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022. CM-Gen: A Neural Framework for Chinese Metaphor Generation with Explicit Context Modelling. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6468–6479, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Longxuan Ma, Weinan Zhang, Shuhan Zhou, Churui Sun, Changxin Ke, and Ting Liu. 2023. I run as fast as a rabbit, can you? A Multilingual Simile Dialogue Dataset. *arXiv preprint*. ArXiv:2306.05672 [cs].

Anthony M Paul. 1970. Figurative language. *Philosophy & Rhetoric*, pages 225–248.

Reenu Paul, Wincy Abraham, and Anitha S Pillai. 2024. Malupama-figurative language identification in malayalam-an experimental study. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 357–367.

Yujie Shao, Xinrong Yao, Xingwei Qu, Chenghua Lin, Shi Wang, Stephen W Huang, Ge Zhang, and Jie Fu. 2024a. Cmdag: A chinese metaphor dataset with annotated grounds as cot for boosting metaphor generation. *arXiv preprint arXiv:2402.13145*.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Zhe Li, Hujun Bao, and Xipeng Qiu. 2024b. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *Science China Information Sciences*, 67(5):152102.

Zhou Shuhan, Ma Longxuan, and Shao Yanqiu. 2023. Exploring Accurate and Generic Simile Knowledge from Pre-trained Language Models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 918–929, Harbin, China. Chinese Information Processing Society of China.

Wei Song, Jingjin Guo, Ruiji Fu, Ting Liu, and Lizhen Liu. 2020. A Knowledge Graph Embedding Approach for Metaphor Processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1–1.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. *arXiv preprint arXiv:2106.01228*.

Chang Su, Shuman Huang, and Yijiang Chen. 2017. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311.

Chang Su, Jia Tian, and Yijiang Chen. 2016. Latent semantic similarity based interpretation of chinese metaphors. *Engineering Applications of Artificial Intelligence*, 48:188–203.

Imke van Heerden and Anil Bas. 2021. Towards figurative language generation in afrikaans. In *Proceedings of the SIGTYP 2021 Workshop on Typology for Cross-Linguistic NLP*.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Xiangpeng Wei, Zhengyuan Liu, and Jun Xie. 2023. Fantastic Expressions and Where to Find Them: Chinese Simile Generation with Multiple Constraints. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 468–486, Toronto, Canada. Association for Computational Linguistics.

Zhiwei Yu and Xiaojun Wan. 2019. How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 861–871.

Jiayi Zhang, Zhi Cui, Xiaoqiang Xia, Yalong Guo, Yanran Li, Chen Wei, and Jianwei Cui. 2021. Writing polishment with simile: Task, dataset and a neural approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14383–14392.

Danning Zheng, Ruihua Song, Tianran Hu, Hao Fu, and Jin Zhou. 2020. "love is as complex as math": Metaphor generation system for social chatbot. In *Chinese Lexical Semantics: 20th Workshop, CLSW 2019, Beijing, China, June 28–30, 2019, Revised Selected Papers 20*, pages 337–347. Springer.

A Dataset

A.1 Tamil Simile Dataset (TSD)

Examples in TSD are shown in Table 4.

Sentence	Tenor	Vehicle	Context
கடல் போல பெரிதாக நீ நின்றாய்.	கடல்	நீ	பெரிதாக
You stood as big as the sea.	sea	you	big
பறவை போலே பறந்து செல்வோம்.	பறவை	நாம்	பறந்து
Let's fly like a bird.	bird	Let's	fly
ஒரு கோயில் போல் இந்த மாளிகை.	கோயில்	மாளிகை	புனிதமானது
This mansion is like a temple.	temple	mansion	sacred
வழியிலே தங்கத்தகடு போல மின்னிய தவளை தத்திச் சென்றது.	தங்கத்தகடு	தவளை	மின்னிய
On the way, a frog that glittered like a gold plate jumped away.	gold plate	frog	glittered

Table 4: Examples of annotated similes in the TSD.

Towards Effective Emotion Analysis in Low-Resource Tamil Texts

Priyatharshan Balachandran[1] **Uthayasanker Thayasivam**[1]
Randil Pushpananda[2] **Ruvan Weerasinghe**[2]
[1]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
[2]University of Colombo School of Computing, University of Colombo, Sri Lanka
balachandran.24@cse.mrt.ac.lk, rtuthaya@cse.mrt.ac.lk
rpn@ucsc.cmb.ac.lk, arw@ucsc.cmb.ac.lk

Abstract

Emotion analysis plays a significant role in understanding human behavior and communication, yet research in Tamil language remains limited. This study focuses on building an emotion classifier for Tamil texts using machine learning (ML) and deep learning (DL), along with creating an emotion-annotated Tamil corpus for Ekman's basic emotions. Our dataset combines publicly available data with re-annotation and translations. Along with traditional ML models we investigated the use of Transfer Learning (TL) with state-of-the-art models, such as BERT and Electra based models. Experiments were conducted on unbalanced and balanced datasets using data augmentation techniques. The results indicate that Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) performed well with TF-IDF and BoW representations, while among Transfer Learning models, LaBSE achieved the highest accuracy (63% balanced, 69% unbalanced), followed by TamilBERT and IndicBERT.

1 Introduction

Emotional Analysis (EA), an extended version of sentiment analysis (SA), extracts emotions from human output using physiological qualities like voice, looks, hand motions, body developments, heartbeat and blood pressure (Chuang and Wu, 2004). R. W. Picard emphasized computers must understand emotions for effective human-computer interaction (Picard, 1997). In this digital era, the divide between ethnic groups and communities has diminished where people love to communicate, understand and experience diversity. Computer language translation plays a prominent role here though it can lead to misinterpretations of emotions within the context on certain occasions. Ze-Jing Chuang and Chung-Hsien Wu's Multi-Modal Emotion Recognition research combining speech and text produced better results than either input alone (Chuang and Wu, 2004). They created the textual model by defining keywords for emotions with emotion modification values. EA remains complex when processing only textual data compared to speech and vision.

To proceed with a structured analysis, selecting a reliable and widely accepted emotional categorization framework is essential. Ekman's basic emotions—Anger, Disgust, Fear, Joy, Sadness, and Surprise caters the above requirement in the research community (Ekman, 1992). Proposed by psychologist Paul Ekman, this scheme was developed based on cross-cultural studies that demonstrated these emotions as universal. Also the mentioned schema found to be consistently recognizable across different societies. The framework has proven highly valuable in emotion recognition tasks for human-computer interaction systems, social robotics, and content analysis.

Creating a well-annotated emotion dataset with Ekman's basic emotions spectrum poses a major challenge for Tamil language EA, as existing datasets exhibit significant class imbalances as well as the emotions not directly aligning with the schema. The TamilEmo dataset (Vasantharajan et al., 2021) requires reclassification with linguistic expert input for Ekman's basic emotions and potential re-annotation to validate the classification. While ACTSEA (Jenarthanan et al., 2019) is not fully publicly available with less than 500 samples, this research aligns with Ekman's basic emotions. The rest of the sentences that do not belong to any of the above is considered Neutral. With current state of the art (SOTA) transformer-based approaches, a proper dataset with class balance will significantly contribute to this research and result in better accuracy than previous works (Vasantharajan et al., 2021; Gokhale et al., 2022). This research aspires to create a balanced Tamil emotion annotated corps and improve emotion detection/recognition by using Natural Language Processing (NLP) with

Machine Learning (ML) and Deep Learning (DL) techniques.

2 Related Works

Emotions can be understood through punctuation, catchphrases, syntax, and semantic data (Chuang and Wu, 2002). The SNoW learning architecture outperformed baseline Naive model and Bag Of Words (BoW) approach (Alm et al., 2005), while Wu et al. presented automatic emotion recognition through semantic labels and attributes (Wu et al., 2006).

A hybrid keyword-based and learning-based approach using SVM achieved 96.43% accuracy (Binali et al., 2010). Shivhare proposed an Ontology method based on commonsense knowledge and interrelationship between entities and core vocabulary (Shivhare and Khethawat, 2012). For Japanese earthquake-related tweets, Vo B and Collier N concluded that simple N-gram features performed best using MNB model (Vo and Collier, 2013).

Canales L and Martinez-Barco's survey discussed computational approaches categorized as lexicon-based and ML-based, noting keyword-based approaches (Strapparava and Mihalcea, 2008), ontology-based (Shivhare and Khethawat, 2012) and statistical approaches (Chuang and Wu, 2002) as lexical methods. Their findings showed keyword-based approaches yield higher accuracy, while supervised learning outperforms unsupervised methods despite requiring resource-intensive annotated datasets (Canales and Martínez-Barco, 2014). SVM has been a traditional supervised learning technique for EA (Hakak et al., 2017), though Nasir A et al. found MNB models perform better than SVM, decision tree algorithm and k-nearest neighbour methods (Ab. Nasir et al., 2020).

2.1 Emerging of Transformers

The introduction of transformers revolutionized the DL field (Vaswani et al., 2017), with BERT becoming SOTA in many NLP implementations despite higher resource consumption (Devlin et al., 2018). Various BERT variants emerged, including mBERT and ALBERT, while XLM-RoBERTa later outperformed mBERT (Conneau et al., 2019b).

Electra emerged as a resource-efficient alternative to BERT (Clark et al., 2020), while Huang C et al.'s ensemble method combining HRLCE and BERT achieved a macro-F1 score of 0.7709 (Huang et al., 2019). Yang K et al. enhanced pre-trained models using MLM and NSP (Yang et al., 2019). Al Omari H, Abdullah M and Shaikh S executed a dual model using BiLSTM and BERT, resulting in an F1 score of 0.748 (Al-Omari et al., 2020). Acheampong F et al.'s review recommends exploring more BERT variants and ensemble models (Acheampong et al., 2021).

Comparative analyses show BERT and Electra outperform RoBERTa, XLM-R and XLNet in fine-grained emotions detection with lower training time (Frye and Wilson, 2022), though Cortiz D found DistillBERT, RoBERTa and XLNet superior to Electra (Cortiz, 2021). Zhang S, Yu H and Zhu G's Electra-based model with attention mechanism and BiLSTM achieved mean accuracies of 94.657 and 93.713 for Chinese language emotion detection (Zhang et al., 2022).

2.2 Research on Tamil Language

Tamil language EA research remains limited compared to other languages. For sentiment analysis, Padmamala R and Prema V's RNN approach achieved 71.1% accuracy (Padmamala and Prema, 2017), while Shanmugavadivel K et al.'s CNN with Bi-LSTM achieved 0.66 accuracy on tamil code-mixed texts (Shanmugavadivel et al., 2022). Sajeetha T's experiments with multiple approaches achieved 79% accuracy using fastText (Thavareesan and Mahesan, 2019), later improving to 88% accuracy using Word2vec and fastText with rule-based approach (Thavareesan and Mahesan, 2020). Sharmista's product review sentiment analysis concluded that ensemble methods produced optimal results (Ramaswami, 2020).

2.3 Research on Tamil Emotion Analysis

Dakshina k and Sridhar R's LDA-based emotion recognition for Tamil songs achieved 72% accuracy using supervised learning with 160 songs and 5 annotators (Dakshina and Sridhar, 2014). Charangan V et al.'s TamilEmo corpus classified 31 emotions from 42,686 sentences scraped from YouTube comments. These samples were annotated with an inter-annotator agreement of 0.7452. A major concern in the dataset is the class imbalance among emotion categories, with the emotion "admiration" having the highest sample count of 6,682 samples, while the emotion "desire" has the lowest sample count, with only 208 samples. Their ML methods achieved a maximum 0.42 F1 score (Vasantharajan et al., 2021). Gokhale O et al. attempted transformer ensemble method and could not achieve

significant improvements for the very same dataset (Gokhale et al., 2022).

Rajalakshmi et al.'s investigation of emoji impact in Tamil Texts showed that replacing emojis with keywords performed best, followed by emoji-present and emoji-removed approaches. Their TF-IDF and XGBoost combination outperformed the MuRIL pre-trained model (Rajalakshmi et al., 2022). This shows that containing the emojis in the dataset is essential for higher results.

3 Dataset Overview

In this study, we adopted Ekman's basic emotions: anger, disgust, fear, joy, sadness, surprise, and neutral since they are widely accepted and frequently used in emotion research, especially in high-resource languages. This approach is a standard in emotion classification tasks, making it a suitable framework to extend to the Tamil language, where similar work has been limited. By aligning with this framework, we aim to standardize emotion classification in Tamil and provide a valuable resource for future research.

3.1 Data Collection

TamilEmo(Vasantharajan et al., 2021) was the only publicly available dataset which was emotion annotated in Tamil language. The TamilEmo dataset had 31 emotion classes and they were grouped into seven primary emotions: Hope, Neutral, Love, Bewilderment, Disgrace, Pathos and Laughter. Of these seven emotions, only three could be mapped directly to Ekman's basic emotions as $Neutral \longrightarrow Neutral$, $Pathos \longrightarrow Sad$ and $Laughter \longrightarrow Joy$. For the other emotions, we had to go with the fine-grained emotions of 31 classes.

Emotions	Angry	Disgust	Fear	Joy	Neutral	Sad	Surprise	Unclassified
admiration	☐	☐	☐	☑	☑	☐	☑	☐
amusement	☐	☐	☐	☑	☑	☐	☐	☐
anger	☑	☐	☐	☐	☐	☐	☐	☐
annoyance	☑	☑	☐	☐	☑	☐	☐	☐
anticipation	☐	☐	☐	☐	☐	☐	☐	☑
approval	☐	☐	☐	☐	☐	☐	☐	☑
caring	☐	☐	☐	☐	☐	☐	☐	☑
confusion	☐	☐	☐	☐	☐	☐	☐	☑
curiosity	☐	☐	☐	☑	☑	☐	☑	☐
desire	☐	☐	☐	☐	☐	☐	☐	☑
disappointment	☑	☐	☐	☐	☑	☑	☐	☐
disapproval	☐	☐	☐	☐	☐	☐	☐	☑
disgust	☐	☑	☐	☐	☐	☐	☐	☐
embarrassment	☐	☑	☐	☐	☑	☑	☐	☐
excitement	☐	☐	☐	☑	☑	☐	☐	☐
fear	☐	☐	☑	☐	☐	☐	☐	☐
gratitude	☐	☐	☐	☑	☑	☐	☑	☐
grief	☐	☐	☐	☐	☐	☑	☐	☐
joy	☐	☐	☐	☑	☐	☐	☐	☐
love	☐	☐	☐	☑	☑	☑	☐	☐
nervousness	☐	☐	☑	☐	☑	☐	☐	☐
neutral	☐	☐	☐	☐	☑	☐	☐	☐
optimism	☐	☐	☐	☑	☑	☐	☑	☐
pride	☐	☐	☐	☑	☑	☐	☑	☐
realization	☐	☐	☐	☐	☐	☐	☐	☑
relief	☐	☐	☐	☑	☐	☐	☐	☐
remorse	☐	☐	☐	☐	☐	☑	☐	☐
sadness	☐	☐	☐	☐	☐	☑	☐	☐
surprise	☐	☐	☐	☐	☐	☐	☑	☐
teasing	☐	☐	☐	☑	☐	☐	☐	☐
trust	☐	☐	☐	☐	☐	☐	☐	☑

Figure 1: Emotion Mapping of TamilEmo Dataset

Under the guidance of our language expert panel, these 31 classes of emotions were mapped to the emotion categories. If any of the emotions were ambiguous and could not be directly mapped, so they were classified into mixed emotions. Few of them couldn't be concluded to any set of emotions or could fall under more than three emotions, so they were categorized as unclassified. All the mixed emotions were included under neutral emotion as well to be sure when annotating. Figure 1 shows how we categorized the emotions. When validating the samples with corresponding emotions, we learned that many samples had been contradictorily annotated in the original study. Another main issue with this dataset is the class imbalance. The emotion admiration has 6682 samples, and the emotion desire has only 208 samples. It was understood that the samples in this dataset would not be sufficient to have a balanced dataset. Therefore we had options to scrape data from the web with keywords and annotate or translate an available English dataset to Tamil and validate them.

We used an English Emotion Dataset (Saravia et al., 2018), which is publicly available in Kaggle and Huggingface as our secondary dataset. This dataset contained English Twitter messages with six basic emotions: anger, fear, joy, love, sadness, and surprise. Except for the emotion of love, all the other emotions directly corresponded to our study. The emotions distributions can be seen in Figure 2

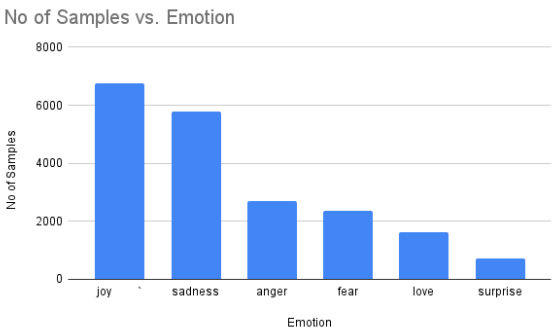


Figure 2: Class Distribution - English Emotion Dataset

3.2 Data Annotation

The datasets were divided into chunks of a maximum of 1000 - 1100 samples to make it easier for the annotators. Every sample was annotated by a pair of annotators using separate spreadsheets without any influence from each other. The annotators were instructed to disregard sarcasm and interpret sentences by their literal meaning since

satire is considered out of scope in our study. The native Tamil-speaking undergraduates of the University of Colombo School of Computing were the annotators.

The samples which were explicitly categorized as in the original study were annotated by selecting whether it is correctly classified or misclassified. For the other emotions, the possible emotions were listed in a drop-down and the annotators were asked to choose the best option. When both annotators completed their annotations, the results were compared, and a third one annotated the contradicting samples. Then the maximum of the emotions selected was made final. In some instances, all three annotators' choices differed from each other. In that case, those samples were filtered out from the final dataset.

The English Emotion dataset was combined as one single dataset CSV file which was initially divided into train, test and validation datasets. Translation was done using Google translate and the annotators were asked to annotate whether the translation made sense or not by selecting either yes or no. Then they were also asked to give points according to the samples giving justice to the emotions. When the pair of annotators had done their work, and checked whether both agreed that the translation was correct and whether the point total was above half of the maximum value. If they contradicted the translation, a third annotator validated those samples. If most of them selected "yes" for the translation, then again as before, the total points were checked for more than half the maximum value. Then the samples were finalized to their corresponding emotions and the rest were abandoned. The point format is as follows.

- 0: Does not align with the emotion.
- 1-4: Have some context related to the emotion, but the translation of the sentence is not appropriate (the overall sentence does not make sense).
- 5: Have context related to the emotion, but the translation is ambiguous, which might exhibit mixed emotions.
- 6-10: Have descent alignment with the emotion.

As this whole annotation process is manual it was a huge concern. The time taken to annotate was longer than anticipated, and it was not easy to manage the annotators. These are the few main challenges we faced during this phase.

- Due to pair wise annotation, the datasets could be annotated at a rate of half the annotators only.
- Inconsistency and slow process of few annotators, where continuous annotation of the next dataset assigned to them was not possible with everyone.
- Have to wait for both the annotators to finish annotating so we can validate it with the third annotator.
- For certain samples, the annotators were not able to conclude their results to a specific emotion. Those samples were finally excluded.

After continuous annotations and validations, at the end approximately more than 50,000 annotations were completed and the final dataset ended with 16804 samples. Figure 3 depicts the final dataset overview.

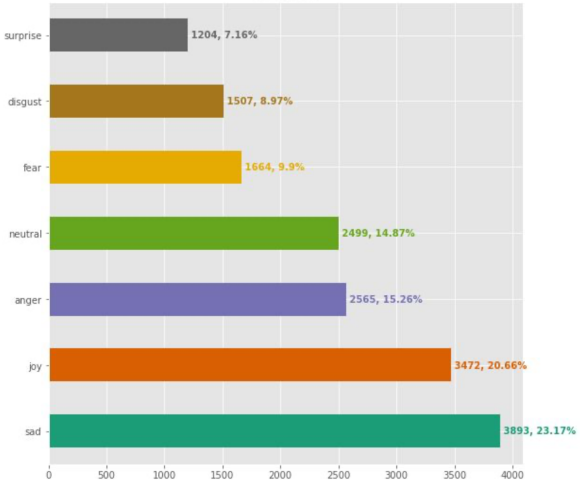


Figure 3: Final Dataset Overview

A balanced dataset could not be generated in the last stage as planned. Several factors contributed to this such as, inherent class imbalance, language-specific challenge, annotation challenges, inherent imbalance in real-world data and time constraints.

The following are the Average Cohen's Kappa values for the annotations of all datasets in Table 1.The average inter-annotator agreement between annotators A and B: 72%, B and C: 69%, A and C: 79% and the average of all inter-annotator agreements is 73%.

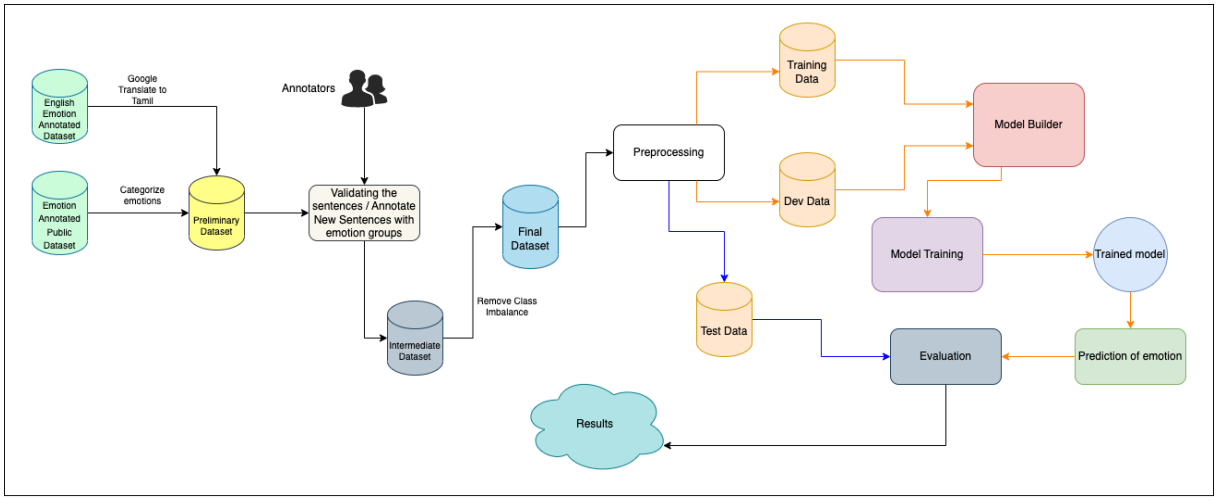


Figure 4: High level approach of the study

	A ↔ B	B ↔ C	A ↔ C	Average
Average Kappa for Annotators	72%	69%	79%	73%

Table 1: Kappa statistics Average for the Final Dataset

4 Experiments and Evaluation

Figure 4 describes the methodology of the research in a higher level. The following sections have detailed information on the experiments and the results.

4.1 Preprocessing

Several measures were taken during the preprocessing phase to ensure the quality and consistency of our dataset. First, the hyperlinks were deleted and profile tags, as well as the white spaces, because they do not contribute to the emotional substance of the text. We also chose against eliminating punctuation and emoticons because they can considerably alter the emotions represented in the samples. Similarly, English terms were not eliminated because they, too, could reflect feelings in some cases.

4.2 Models Utilized

Both ML and DL approaches were employed in the study. In addition to that dual model approach of involving pre-trained models and deep learning models were also experimented. In the traditional ML experiments SVM, random forest, naive bayes, multinomial naive bayes (MNB), decision tree, passive aggressive and K-Nearest Neighbour (KNN) were included. In the other hand for DL models, Universal Sentence Encoder CMLM - Multilingual Base (Yang et al., 2021), MuRIL-Large (Multilingual Representations for Indian Languages) (Khanuja et al., 2021), BERT Multilingual Cased (bert_multi_cased_L-12_H-768_A-11) (Devlin et al., 2018), XLM-R0BERTa Multilingual Cased (xlm_roberta_multi_cased_L-24_H-1024_A-16) (Conneau et al., 2019a), DistilBERT (distilbert_multi_cased_L-6_H-768_A-12) (Sanh et al., 2019), Tamillion (An Electra based monolingual Tamil pre-trained model), IndicBERT, LaBSE (Language-agnostic BERT Sentence Encoder) (Kakwani et al., 2020), LaBSE (Language-agnostic BERT Sentence Encoder) (Feng et al., 2020), TamilBERT(Joshi, 2022) were utilized. Then as mentioned for dual model approach, CNN with Bi-LSTM, CNN with Bi-GRU and LaBSE with Bi-LSTM were combined. The architecture and training source details of the models are attached in the Appendix A

4.3 Unbalanced Final Dataset

When rendering the results, the best-performing models in this experiment appeared to be SVM and Passive Aggressive, especially when combined with TF-IDF Unigram with Bigram and FastText text representations. The performance of Naive Bayes and Multinomial Naive Bayes models varied significantly depending on the text representation used. They performed better with Bag of Words Unigram and Bigram representations, consistent with the previous experiments. However, their performance dropped when using other text representations like the Bag of Words Trigram and TF-IDF Unigram. This is illustrated in Table 2, where the best-performing text representations for each model are highlighted in yellow, and the highest-performing models for each text representation is

indicated in bold.

Decision Tree, K Nearest Neighbor, and Random Forest models showed modest improvements in accuracy in the unbalanced dataset experiment compared to their performance in the previous downsampled dataset experiment. These models might have been less sensitive to class imbalance, and their performance could have depended more on the quality and quantity of the available data.

	Naive Bayes	MNB	SVM	Decision Tree	KNN	Random Forest	Passive Aggressive
Bag of Words							
Unigram	0.16	**0.53**	**0.52**	0.44	0.39	**0.51**	0.44
Bigram	0.36	**0.49**	**0.50**	0.45	0.31	**0.48**	0.47
Trigram	0.34	0.39	**0.43**	0.36	-	0.37	0.36
TF-IDF							
Unigram	0.30	0.34	**0.44**	0.34	0.20	0.37	**0.44**
Unigram with Bigram	0.36	0.47	**0.58**	0.45	0.24	0.53	0.54
FastText	0.40	-	**0.56**	0.37	0.44	0.53	**0.56**

Table 2: Accuracy comparison of ML Models for the Final Unbalanced Dataset

In terms of accuracy, most models showed an increase in performance when using the unbalanced dataset compared to the downsampled one. This might have been due to the larger amount of data available for training, which generally helps the models better understand the patterns and capture more nuanced relationships between the features and target emotions. Moreover, the class imbalance in the unbalanced dataset might have also played a role in the increased accuracies since the models were now better exposed to the majority class, which is more frequently seen in real-life scenarios.

Model	Accuracy
Universal Sentence Encoder	61%
MuRIL-Large	60%
BERT Multilingual Cased	51%
XLM-R0BERTa Multilingual Cased	43%
Distil-BERT	49%
Tamillion	62%
IndicBERT	64%
TamilBERT	67%
LaBSE	**69%**

Table 3: Accuracy comparison of Transformer Models for the Final Unbalanced Dataset

When comparing models focused on one language (monolingual) and models that worked with multiple languages (multilingual), we found that TamilBERT and Tamillion (monolingual models) had higher accuracies. This could have been because they were designed specifically for Tamil. However, LaBSE, a multilingual model, also performed very well, achieving an accuracy of 69%.

Among Indian language-based models, IndicBERT, which was trained on 12 major Indian languages, achieved an accuracy of 64%. This indicates that a model trained on multiple Indian languages can still perform well for Tamil language classification. When examining different model types, BERT-based models like TamilBERT, LaBSE, and BERT Multilingual Cased showed varying levels of success. TamilBERT and LaBSE performed significantly better. The ELECTRA-based model, Tamillion, also produced good results with an accuracy of 62%. As mentioned in the literature, ELECTRA models tend to perform well, but we were unable to find more pre-trained ELECTRA models related to Tamil for experimentation. Finally, the ALBERT-based model, IndicBERT, demonstrated strong performance with an accuracy of 64%.

In the current experiment, we observed a significant improvement in the accuracies of the transfer learning models compared to the machine learning models on the unbalanced dataset. LaBSE and TamilBERT achieved higher accuracies of 69% and 67% respectively, which were considererably higher than the best machine learning model using TF-IDF Unigram with Bigram (58%). This highlights the advantages of utilizing pre-trained models.

Model	Accuracy
CNN with Bi-LSTM	56%
CNN with Bi-GRU	56%
LaBSE with Bi-LSTM	**61%**

Table 4: Accuracy comparison of Dual Models for the Final Unbalanced Dataset

Upon examining the outcomes of the hybrid model approach applied to the same dataset, it became evident that their effectiveness lay somewhere between the machine learning and transfer learning models. The CNN combined with a Bi-LSTM model reached an accuracy of 56%, while the CNN paired with a Bi-GRU model achieved a 56% accuracy rate. The LaBSE, alongside a Bi-LSTM model incorporating transfer learning, produced a higher accuracy level of 61%. Despite the enhancements displayed by the hybrid models compared to the machine learning models, they failed to outperform transfer learning models like the fine-tuned models.

4.4 Balanced Final Dataset

To achieve a balanced dataset, a decision was made to split the original samples into training and test sets before proceeding with data augmentation. The split involved allocating 15% of the number of samples in the minority class (241 samples) as the test data, while the remaining samples were retained for training. It was determined that 2250 samples per emotion would be used for the training data.

Referring to the work of Jie and Gao (Gao, 2020) on Data Augmentation in Solving Data Imbalance Problems, it was found that translation proved to be an effective technique for upsampling textual data. Based on this finding, translation was chosen as the upsampling technique. Google Translate was utilized for the translation process. Initially, the dataset was translated from Tamil to English and then back to Tamil. Following the translation, a careful selection process was implemented to ensure that the majority of the samples remained original, with only the necessary number of samples required for balancing the dataset being included from the translated samples. These samples were randomly selected from each class within the translated set.

	MNB	SVM	Random Forest	Passive Aggressive
Bag of Words				
Unigram	**0.47**	**0.46**	0.45	0.39
Bigram	**0.45**	0.43	0.43	0.41
TF-IDF				
Unigram	**0.40**	**0.41**	0.34	**0.40**
Unigram with Bigram	**0.48**	**0.48**	**0.48**	0.44
FastText	-	**0.47**	0.44	0.44

Table 5: Accuracy comparison of ML Models for the Final Balanced Dataset

The results of this experiment indicate that the models performed below expectations compared to the machine learning experiments conducted on the unbalanced dataset, with the exception of Multinomial Naive Bayes combined with TF-IDF Unigram with Bigram text representation, which showed a slight improvement. When compared to the previous balanced dataset experiment, the differences in results were relatively minor. The models performed better overall compared to the earlier experiments, except for SVM, Random Forest, and Passive Aggressive with the text representations TF-IDF Unigram with Bigram and FastText. However, when compared to the deep learning (DL) models, the performance of these models fell significantly below the average.

In this experiment, when analyzing the results exclusively, TF-IDF Unigram with Bigram emerged as the top-performing text representation, achieving accuracy ranging from 44% to 48%. Similarly, Multinomial Naive Bayes stood out among the models, also with accuracy ranging from 44% to 48%. The Bag of Words (BoW) Unigram and SVM model provided tough competition to the top-ranking models.

Model	**Accuracy**
CNN with Bi-LSTM	52%
CNN with Bi-GRU	50%
LaBSE with Bi-LSTM	55%
Tamillion	57%
TamilBERT	62%
LaBSE	**63%**

Table 6: Accuracy comparison of Transformer and Dual Models for the Final Balanced Dataset

As expected, the deep learning (DL) models performed better than the machine learning (ML) models. Among the hybrid models, LaBSE with Bi-LSTM achieved an accuracy of 55%, which was higher than the other models. CNN with Bi-LSTM outperformed CNN with Bi-GRU, while in the unbalanced dataset, they performed at similar levels. The transfer learning model, Tamillion, achieved an accuracy of 57%, which was relatively lower than the bert-based models LaBSE and TamilBERT. Among the fine-tuned combinations, LaBSE achieved the highest accuracy of 63%. Appendix B describes this best performing model architecture. It is worth noting that these transfer learning models outperformed some of the other transfer learning models from the unbalanced dataset, even though the accuracy of all the models for this balanced dataset dropped compared to the previous experiment.

4.5 Error Analysis

For each model outputs, we did observation study on the confusion matrix followed by the error analysis. The observation pointed out that the confusion matrices produced throughout the experiment were quite similar, which suggests that the dataset maintains internal consistency. The Confusion matrix of the best model LaBSE is displayed in Figure 5. When observing the confusion matrix, commonly Disgust is confused with Anger and Joy with Neutral even with the balanced dataset. 105 samples

of Disgust have been falsely predicted as Anger and 57 vice versa. 48 samples of Joy have been incorrectly predicted as Neutral, and 41 samples in the other way. Except for Fear and Sadness, other emotions have some confusion with Neutral emotion, which is obvious that these emotions might also tend to be neutral on certain occasions.

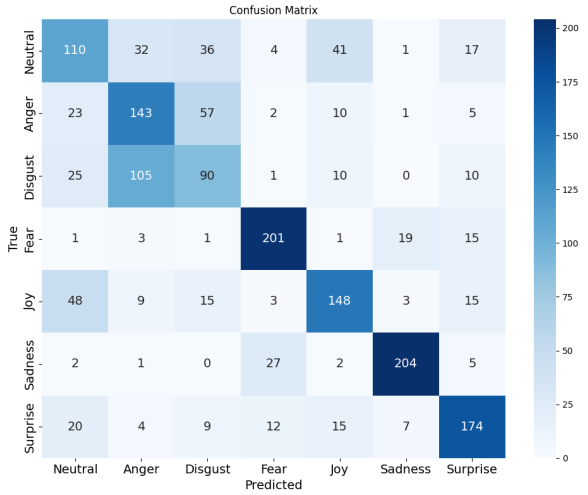


Figure 5: Confusion Matrix of LaBSE model for the Final Balanced Dataset

The most common thing observed throughout this experiment is that the emotion Disgust is being confused with Anger and Neutral with Joy. Since the emotion Disgust is the minority class with a low number of 241 samples, and the confused number of samples is significant and has a huge impact. So an error analysis was done using the LIME library for the emotion Disgust, which was predicted as Anger, and the analysis is attached in Appendix C. As per the analysis result, the model predicted this as anger with 97% confidence and showed Disgust as 0.02% only. So when we look at this example, the sentence can also actually be said as it represents Anger because of the presence of words expressing 'killing' and 'cursing'. This particular sentence has some words relating to castism and extremism based on caste, which might have led the annotators to annotate it as Disgust which also makes sense. Because the samples in Disgust or Anger get confused with each other, they may be exhibiting both emotions in a way, these confusions may have occurred, and the results are affected correspondingly.

5 Discussion

Throughout the study, various text representation techniques have been utilized. TF-IDF Unigram with Bigram yielded higher results in almost all our experiments. Other than these FastText pre-trained models also gave fairly competitive results to the above representations. Considering the transfer learning models, their own preprocessor should be used to achieve better results.

Out of the models that were trained, SVM scored the best and consistently gave better results throughout the ML experiments, especially combined with TF-IDF Unigrams with Bigrams. MNB, Random Forest and Passive Aggressive can be considered alternative models with slightly below-par performance.

Under transfer learning models, the LaBSE model is the highest achieving model with an accuracy 69% for the unbalanced and 63% for the balanced dataset after fine-tuning. Along with that, TamilBERT and IndicBERT also gave fair results. The Electra-based Tamillion model did not perform as expected from the literature (Zhang et al., 2022), where Electra models gave a nearly par performance value with the BERT models. Finally, considering the hybrid models, LaBSE combined with Bi-LSTM models has served better in the hybrid category. Out of all the model categories, transfer learning approaches outperformed every other category.

We compared our results with the original study, where our results can be compared with their **7-class group results**. So far in this experiment, **TF-IDF with Unigrams and Bigrams** combining the **MNB model** for our final balanced dataset of **15,750 samples** has performed **F1-Score of 0.46**, which is higher than the original study. The size of the dataset also matters when comparing the results, as well as the class distribution. The smaller dataset might also be a reason for us reaching higher results. Since the original study employs an **unbalanced dataset**, it is fair to compare the results of our unbalanced dataset. Our best-performing model, **LaBSE** scored an **F1-Score of 0.64** while **TamilBERT**, **IndicBERT**, and **Tamillion** scored **0.62**, **0.60**, and **0.57**, respectively, which are **significantly better results than the original study**.

6 Limiations and Future work

The limitations include the scarcity of quality emotion-annotated datasets for the Tamil language, especially for Ekman's basic emotions. Along with that, the fine-grained nature of the available dataset which did not align with Ekman's basic emotions

also led to complete re-annotation. Additionally, due to class imbalance, the emotion with the highest sample count had to be reduced to 2,250, enabling decent augmentation for the lowest sample count, which resulted in an overall reduction in total sample count.

As a continuation of this study, there is a vast scope to experiment with rule-based preprocessing where negation words and word polarity can be considered. This can be extended to code mixed corpus as well. Investigating more combinations of hybrid models and ensemble approaches with the trained models might give better results. On top of this contrastive learning as well employment of large language models leveraging prompt engineering can be considered. Identifying sarcasm is another dimension to explore in this domain. This textual classification can be integrated with speech-to-text jobs and evolve into an emotion classifier for speech. Other than Ekman's basic emotions, the writing styles such as formal, casual etc. can also be classified.

Acknowledgment

A heartfelt thanks to the authors of "TamilEmo: Finegrained Emotion Detection Dataset for Tamil for providing us with their dataset to conduct my research. Our sincere gratitude for my advisor and language expert Mr V. Vimalathithan for his invaluable guidance and feedback. Special thanks go to the native Tamil students of University of Colombo School of Computing who took part in annotating the dataset.

References

Ahmad Fakhri Ab. Nasir, Eng Nee, Chun Sern Choong, Ahmad shahrizan Abdul ghani, Anwar P P Abdul Majeed, Asrul Adam, and Mhd Furqan. 2020. Text-based emotion prediction system using machine learning approach. *IOP Conference Series: Materials Science and Engineering*, 769:012022.

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54:5789–5829.

Hani Al-Omari, Malak A. Abdullah, and Samira Shaikh. 2020. Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. pages 226–232. Institute of Electrical and Electronics Engineers Inc.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction.

Haji Binali, Chen Wu, and Vidyasagar Potdar. 2010. Computational approaches for emotion detection in text.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey.

Ze-Jing Chuang and Chung-Hsien Wu. 2002. Emotion recognition from textual input using an emotional semantic network.

Ze-Jing Chuang and Chung-Hsien Wu. 2004. Multi-modal emotion recognition from speech and text.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Unsupervised cross-lingual representation learning at scale.

Diogo Cortiz. 2021. Exploring transformers in emotion recognition: a comparison of bert,distillbert, roberta,xlnetandelectra.

K. Dakshina and Rajeswari Sridhar. 2014. Lda based emotion recognition from lyrics. volume 27, pages 187–194. Springer Science and Business Media Deutschland GmbH.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Paul Ekman. 1992. Are there basic emotions?

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Robert H Frye and David C Wilson. 2022. Comparative analysis of transformers to support fine-grained emotion detection in short-text data.

JIE Gao. 2020. Data augmentation in solving data imbalance problems.

Omkar Gokhale, Shantanu Patankar, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. Optimize$_prime$@$dravidianlangtech - acl2022 : Emotionanalysisintamil$.

Nida Manzoor Hakak, Mohsin Mohd, Mahira Kirmani, and Mudasir Mohd. 2017. Emotion analysis: A survey. In *2017 international conference on computer, communications and electronics (COMPTELIX)*, pages 397–402. IEEE.

Chenyang Huang, Amine Trabelsi, and Osmar R. Zaïane. 2019. Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert.

Rajenthiran Jenarthanan, Yasas Senarath, and Uthayasanker Thayasivam. 2019. *ACTSEA: Annotated Corpus for Tamil Sinhala Emotion Analysis*.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

R. Padmamala and V. Prema. 2017. Sentiment analysis of online tamil contents using recursive neural network models approach for tamil language. In *2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, pages 28–31.

R W Picard. 1997. Affective computing.

Ratnavel Rajalakshmi, Faerie Mattins R, Srivarshan Selvaraj, Antonette Shibani, Anand Kumar M, and Bharathi Raja Chakravarthi. 2022. Understanding the role of emojis for emotion detection in Tamil. In *Proceedings of the First Workshop on Multimodal Machine Learning in Low-resource Languages*, pages 9–17, IIIT Delhi, New Delhi, India. Association for Computational Linguistics.

M Ramaswami. 2020. Sentiment analysis on tamil reviews as products in social media using machine learning techniques: A novel study doctor of philosophy in computer science.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech Language*, 76:101407.

Shiv Naresh Shivhare and Prof Saritha Khethawat. 2012. Emotion detection from text.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. *Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation*.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020. *Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts*.

Charangan Vasantharajan, Sean Benhur, Prasanna Kumar Kumarasen, Rahul Ponnusamy, Sathiyaraj Thangasamy, Ruba Priyadharshini, Thenmozhi Durairaj, Kanchana Sivanraju, Anbukkarasi Sampath, Bharathi Raja Chakravarthi, and John Phillip McCrae. 2021. Tamilemo: Finegrained emotion detection dataset for tamil.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Bao-Khanh H Vo and Nigel Collier. 2013. Twitter emotion analysis in earthquake situations.

Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. 2006. Emotion recognition from text using semantic labels and separable mixture models.

Kisu Yang, Dongyub Lee, Taesun Whang, Seolhwa Lee, and Heuiseok Lim. 2019. Emotionx-ku: Bert-max based contextual emotion classifier.

Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2021. Universal sentence representation learning with conditional masked language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6216–6228, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shunxiang Zhang, Hongbin Yu, and Guangli Zhu. 2022. An emotional classification method of chinese short comment text based on electra. *Connection Science*, 34:254–273.

A Overview of the Transfer Learning Models Used

Table 7: Overview of the Transfer Learning Models

Model	Language Type	Architecture	Preprocessor
Universal Sentence Encoder CMLM - Multilingual Base (Yang et al., 2021)	Universal sentence encoder for 100+ languages trained with a conditional masked language model.	The base model employs a 12-layer BERT transformer architecture.	universal-sentenc e-encoder-cmlm/ multi lingual-preprocess
MuRIL-Large (Multilingual Representations for Indian Languages) (Khanuja et al., 2021)	Pre-trained on 17 Indian languages, and their transliterated counterparts.	A BERT Large (24L) model	MuRIL_pre process
BERT Multilingual Cased(Devlin et al., 2018)	Multilingual	BERT architecture. Uses L=12 hidden layers, a hidden size of H=768, and A=12 attention heads	bert_multi_case d_preprocess
XLM-R0BERTa Multilingual Cased(Conneau et al., 2019a)	Multilingual	Uses L=24 hidden layers, a hidden size of H=1024, and A=16 attention heads	xlm_roberta_mu lti_cased_pre process
Distil-BERT (Sanh et al., 2019)	Multilingual	Uses L=6 hidden layers, a hidden size of H=768, and A=12 attention heads	distilbert_multi_ cased_pre process
Tamillion	Monolingual (Tamil)	Model trained with Google Research's ELECTRA	ElectraTokenizer from transformers library
IndicBERT (Kakwani et al., 2020)	Multilingual. Pretrained exclusively on 12 major Indian languages	ALBERT (A Lite BERT for Self-supervised Learning of Language Representations) based model	AlbertTokenizer from transformers library
LaBSE (Language-agnostic BERT Sentence Encoder) (Feng et al., 2020)	Trained for sentence embedding for 109 languages	Based on the BERT architecture and uses a Siamese network with shared weights to learn a joint embedding space for different languages	BertTokenizer from transformers library **or** universal-sentenc e-encoder-cmlm/ multi lingual-preprocess
TamilBERT (Joshi, 2022)	Monolingual (Tamil)	Based on the BERT architecture	BertTokenizer from transformers library

B Best Model Architecture Details

The best performing model architecture consists of the Language-agnostic BERT Sentence Embedding (LaBSE) as the base model with a custom classification head. The complete architecture and training configuration are detailed below.

B.1 Model Architecture

- Input Layer: Text input layer accepting string data
- Base Model:
  - LaBSE Preprocessor
  - LaBSE Encoder
- Classification Head:
  - Dropout (rate = 0.2)
  - Dense Layer (128 units, ReLU activation)
  - Dropout (rate = 0.3)
  - Dense Layer (64 units, ReLU activation)
  - Dropout (rate = 0.1)
  - Output Layer (7 units, Softmax activation)

B.2 Training Configuration

- Optimizer: Adam
- Loss Function: Categorical Cross Entropy
- Early Stopping:
  - Monitor: Validation Loss
  - Training Duration: Stopped at epoch 5

B.3 Model Parameters

- Total Trainable Parameters: 109M*
- Base Model:
  - LaBSE Parameters: 109M*
- Classification Head:
  - Dense Layer 1: $128 \times \text{hidden\_size} + 128$ parameters
  - Dense Layer 2: $64 \times 128 + 64$ parameters
  - Output Layer: $7 \times 64 + 7$ parameters

C Error Analysis Using LIME

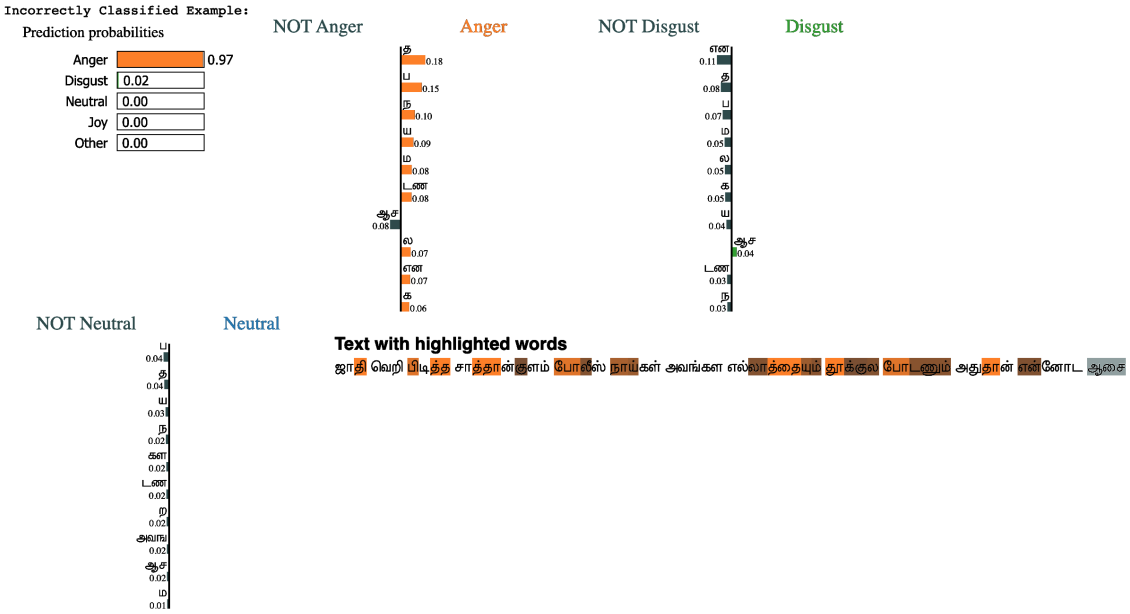


Figure 6: LIME Error Analysis - Disgust Predicted as Anger

Bridging Linguistic Complexity: Sentiment Analysis of Tamil Code-Mixed Text Using Meta-Model

Anusha M D[1], Deepthi Vikram [1], Parameshwar R Hegde[1],

[1]Department of Computer Science,
Yenepoya Institute of Arts Science Commerce and Management,
Yenepoya (Deemed to be University),Balmata, Mangalore

Correspondence: param1000@yahoo.com

Abstract

Sentiment analysis in code-mixed languages poses significant challenges due to the complex nature of mixed-language text. This study explores sentiment analysis on Tamil code-mixed text using deep learning models such as Long Short-Term Memory (LSTM), hybrid models like Convolutional Neural Network (CNN) + Gated Recurrent Unit (GRU) and LSTM + GRU, along with meta-models including Logistic Regression, Random Forest, and Decision Tree. The LSTM+GRU hybrid model achieved an accuracy of 0.31, while the CNN+GRU hybrid model reached 0.28. The Random Forest meta-model demonstrated exceptional performance on the development set with an accuracy of 0.99. However, its performance dropped significantly on the test set, achieving an accuracy of 0.1333. The study results emphasize the potential of meta-model-based classification for improving performance in NLP tasks.

Keywords: Code-mixed, Dravidian Languages, Multi-class, Meta-model, Sentiment Analysis

1 Introduction

In recent years, the analysis of data from social networks and microblogging platforms has garnered significant attention. These platforms are widely used for discussions on a variety of topics, ranging from daily activities and plans to feedback on services and products(Bouazizi and Ohtsuki, 2016). Consequently, businesses and organizations are leveraging such data to extract valuable insights, including user interest in specific topics, satisfaction levels with products and services, and even their intentions and expectations concerning upcoming events like elections or sports competitions.

Another prominent area of research focuses on identifying the attitudes or opinions expressed by users in their posts on specific topics, a process known as "sentiment analysis"(Liu, 2022). The distinctive features of Twitter, a popular microblogging site, make it perfect for sentiment analysis (Memiş et al., 2024). Twitter users can follow others unilaterally, which makes information gathering easier than on many other social networks that require reciprocal connections. It is especially useful for sentiment analysis because of its open format, 140 character limit, and heavy hashtag usage(Bouazizi and Ohtsuki, 2017). Because of the character limit, posts are kept concise and targeted, making it simple to extract insights. Hashtags make Twitter a valuable tool for sentiment analysis by assisting companies in keeping an eye on tweets about their goods or services.

This study investigates hybrid and meta-model techniques for sentiment analysis of Tamil code-mixed text, focusing on multi-class classification. Key contributions include:

- Development of hybrid models (LSTM+GRU, CNN+GRU) for handling code-mixed sentiment data
- Implementation of meta-models (Random Forest, Logistic Regression, Decision Tree) to improve classification accuracy
- Performance evaluation, emphasizing Random Forest's 0.99 accuracy on the development set and generalization issues on the test set

2 Literature Review

This literature review examined sentiment analysis using hybrid models that integrate deep learning architectures to enhance feature extraction and classification, along with meta-models that optimize performance through ensemble methodologies.

2.1 Sentiment Analysis with Hybrid Model

Sentiment analysis plays a critical role in understanding public opinions, particularly from social

media data (Liu, 2022). However, the challenges posed by diverse linguistic structures, code-mixed text, and imbalanced datasets have necessitated the development of innovative hybrid models and optimization techniques.

Tan et al. (2022) addressed these challenges by proposing a hybrid model, RoBERTa-LSTM, which integrates the powerful text encoding capabilities of the pre-trained RoBERTa architecture with the ability of LSTM to capture long-term dependencies. To further improve the model's performance, data augmentation techniques utilizing GloVe embeddings were applied to balance the datasets by oversampling underrepresented classes. This hybrid approach achieved impressive F1-scores of 93%, 91%, and 90% on the IMDb, Twitter US Airline Sentiment, and Sentiment 140 datasets, respectively, highlighting its effectiveness in processing diverse textual data In their exploration of hybrid approaches.

A Bi-LSTM-GRU model combined with a Fuzzy Emotion Extractor (FEE) and the Enhanced Aquila Optimizer (EAQ) proposed by Sherin et al. (2024). This approach improved accuracy by 5.6%, 9.8%, and 8.3% on Sentiment 140, T4SA, and Airline Twitter datasets, respectively. However, limitations like few emotion categories and scalability issues were addressed by incorporating BERT and other optimization techniques

2.2 Sentiment Analysis with Meta-Models

Meta-models, commonly used in stacking ensembles, combine predictions from multiple base models to enhance overall performance (Mekala et al., 2020). In stacking, a meta-model is trained on the outputs of base models to determine the best way to integrate them for more accurate predictions.

Historically, multi-class text classification, especially in sentiment analysis, relied on traditional machine learning methods like Naive Bayes, SVM, and Random Forests. These models used feature extraction techniques such as bag-of-words, TF-IDF, and n-grams but faced challenges in capturing complex semantics and managing large-scale data (Yenter and Verma, 2017). Additionally, Support Vector Regression (SVR) was occasionally applied for text similarity, though its utility was limited when handling diverse text sources (Li et al., 2014). In Yenter and Verma (2017), sentiment analysis was performed using the Doc2Vec embedding model with seven classifiers, including KNN, AdaBoost, and SVM, on the US airline services dataset. The AdaBoost classifier achieved 84.5% accuracy, but the small dataset raised concerns about potential underfitting. Jiang et al. (2019) used Word2Vec and GloVe embeddings with LSTM networks for sentiment analysis on US airline tweets, achieving 75% accuracy in classifying sentiments into positive, neutral, and negative categories.

3 Methodology

The proposed study includes text preprocessing, tokenization, and padding to prepare the data for effective analysis and model training.

3.1 Pre-processing

To improve the quality of the text data, preprocessing techniques were used, such as eliminating user mentions, Tamil stopwords, punctuation, and numerical values. For multiclass classification tasks, the text was tokenized into words and sentiment labels (positive, negative, and neutral) were transformed into binary vectors using one-hot encoding.

3.2 Feature Extraction

Words were converted into numerical indices and padding sequences to 100 tokens using a Keras tokenizer. Word vectors for feature extraction were supplied by Tamil FastText embeddings; zero vectors were allocated to missing words, and the word vectors from a pre-trained model formed an embedding matrix. These procedures set up the data so that machine learning models could process it efficiently for sentiment analysis.

3.3 Model Building

Three models were developed for multi-class classification and integrated through a hybrid approach with a meta-model to enhance performance, as shown in Figures 1 and 2. The hybrid model combines individual strengths for a more robust solution to classification challenges.

Hybrid Models

- LSTM: a type of RNN(Graves and Graves, 2012), model sequences and captures long-term dependencies, making it effective for text data. The model used a pre-trained embedding layer, followed by an LSTM layer with 128 units. To prevent overfitting, dropout layers were added, along with a dense layer of 64 units and ReLU activation. The final output

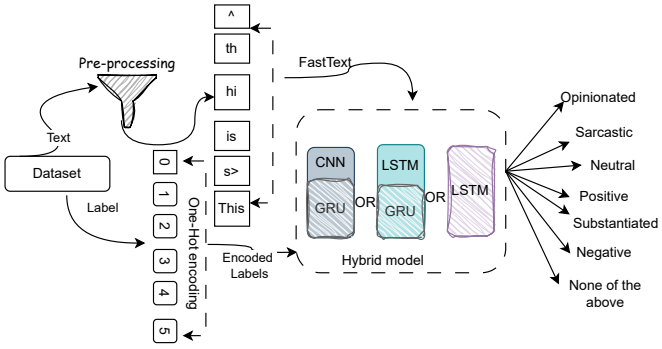


Figure 1: Framework of the Hybrid Sentiment Analysis Model

layer had 7 units with softmax activation for multi-class classification.

- LSTM+GRU Hybrid: combines the strengths of both architectures, with the LSTM layer followed by a GRU layer to capture different aspects of sequential data (MARCELLINA, 2022). GRU units, similar to LSTM, have a simplified structure and fewer parameters. The model uses a pre-trained embedding layer, followed by an LSTM layer (128 units) and a GRU layer (64 units). A dropout layer (rate 0.2) regularizes the network, and the final dense layer (64 units, ReLU activation) is followed by a softmax output layer for class prediction.

- CNN+GRU Hybrid: allows the model to learn both spatial and temporal features by combining CNNs for local feature extraction with GRU layers for processing sequential dependencies (Wu et al., 2020). A max-pooling layer, a 1D convolutional layer (128 filters, kernel size 5), and an embedding layer come first in the model. Sequential dependencies are captured by a 128-unit GRU layer, and the network is regularized by a dropout layer. Finally, a dense layer (7 units, softmax activation) is applied to the output.

Meta-Models(Stacking Ensembles)

To enhance classification performance, a meta-modeling approach was employed, where predictions from base models serve as inputs to a higher-level model that combines these predictions for the final decision (Mekala et al., 2020). This study tested three distinct meta-models:

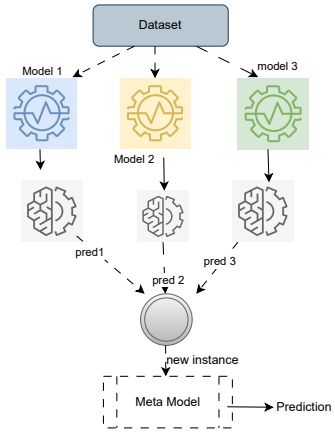


Figure 2: Framework of the proposed Meta-Model, consisting of three models. `pred1`, `pred2`, and `pred3` represent the predictions generated by the three individual models in the hybrid system. The final output is obtained by aggregating these predictions that is new instance.

- Logistic Regression Meta-Model:A linear model that integrates the predictions from base models into a single vector and learns the decision boundary (Taha and Malebary, 2022). Logistic Regression was trained on the stacked predictions from the LSTM+GRU, CNN+GRU, and standalone LSTM models to generate the final classification outcomes.

- Random Forest Meta-Model: An ensemble technique that aggregates the predictions of multiple decision trees (Bjerre et al., 2022). The predictions from the LSTM+GRU, CNN+GRU, and standalone LSTM models served as features for a Random Forest classifier, which was then trained to produce the final sentiment prediction.

- Decision Tree Meta-Model:A tree-based model that makes decisions using input features (Pavel and Soares, 2002). The stacked predictions from LSTM+GRU, CNN+GRU, and standalone LSTM models were used as input to the Decision Tree classifier.

The performance of all three meta-models was evaluated using standard classification metrics includes accuracy, precision, recall, and F1-score. The Random Forest Classifier outperformed the others, delivering the best results on the classification task. The implementation code is available on GitHub

4 Experimental Results

4.1 Dataset

The dataset (Chakravarthi et al., 2025) used in this study is derived from the Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments, shared during the DravidianLangTech@NAACL2025 competition, as outlined in DravidianLangTech@NAACL2025. Specifically curated for multi-class classification tasks, it includes Tamil social media comments with a range of sentiment categories. The dataset is divided into training (4,352 samples) and testing (544 samples) subsets. The sentiment labels span various emotional expressions, and a summary of the label distribution for both the training and testing datasets is presented in Table 1.

Label	Train Count	Test Count
Opinionated	1,361	153
Sarcastic	790	115
Neutral	637	84
Positive	575	69
Substantiated	412	52
Negative	406	51
None of the Above	17	20

Table 1: Label Distribution in Training and Testing Datasets

4.2 Classification

This study experimented with three meta-models (Random Forest, Logistic Regression, and Decision Tree) and three hybrid deep learning models (LSTM, LSTM+GRU, and CNN+GRU) to evaluate sentiment analysis performance on Tamil code-mixed text, focusing on both development and test sets.

The Random Forest meta-model, GRU Hybrid, and CNN+GRU Hybrid models performed best on the development set, leading to their submission in the DravidianLangTech@NAACL 2025 competition. The Random Forest model achieved an impressive 0.99 Macro Average accuracy on the development set but struggled on the test set, with accuracy dropping to 0.1333, suggesting overfitting. The LSTM+GRU and CNN+GRU Hybrid models recorded accuracies of 0.31 and 0.28, respectively, showing potential but not surpassing the Random Forest meta-model.

Class Label	Precision	Recall	F1-Score
Meta Model			
Opinionated	1.00	0.98	0.99
Sarcastic	1.00	0.96	0.98
Neutral	1.00	1.00	1.00
Positive	0.99	0.99	0.99
Substantiated	0.99	1.00	0.99
Negative	0.97	1.00	0.98
None of the Above	1.00	0.98	0.99
Accuracy		**0.99**	
Macro Avg	0.99	0.99	0.99
Weighted Avg	0.99	0.99	0.99
LSTM+GRU Hybrid Model			
Opinionated	0.00	0.00	0.00
Sarcastic	0.14	0.06	0.08
Neutral	0.47	0.80	0.59
Positive	0.32	0.59	0.42
Substantiated	0.26	0.33	0.29
Negative	0.31	0.30	0.30
None of the Above	0.00	0.00	0.00
Accuracy		**0.31**	
Macro Avg	0.22	0.30	0.24
Weighted Avg	0.23	0.31	0.25
CNN + GRU Hybrid Model			
Opinionated	0.02	0.12	0.03
Sarcastic	0.11	0.21	0.14
Neutral	0.80	0.43	0.56
Positive	0.61	0.30	0.40
Substantiated	0.14	0.19	0.16
Negative	0.19	0.24	0.21
None of the Above	0.04	0.50	0.07
Accuracy		**0.28**	
Macro Avg	0.27	0.29	0.23
Weighted Avg	0.45	0.28	0.33

Table 2: Performance metrics for Meta-Model, GRU Hybrid, and CNN + GRU Hybrid.

4.3 Analysis

The literature review indicates that meta-learning techniques are underutilized in multi-class classification, presenting an opportunity for innovation in this area. This study also introduces a novel dataset, publicly available for the first time, which is a significant step in addressing political content classification challenges. By utilizing meta-learning approaches in a new research area, this study expands

meta-learning applications and paves the way for future advancements in multi-class classification tasks.

The dataset exhibited class imbalance, leading to overfitting, where models learned biased patterns toward majority classes. This resulted in a sharp decline in accuracy from the development set to the test set. Additionally, while hybrid models are effective, they complicate computational processes. The study also overlooks transformer-based models, which have demonstrated potential for enhancing performance in similar tasks. To better investigate larger datasets, tackle class imbalance, and examine advanced architectures, further research is needed.

5 Conclusion

This study used hybrid deep learning models and sophisticated feature extraction techniques to examine sentiment analysis on Tamil code-mixed text. Due to its ability to capture temporal and spatial dependencies, the LSTM+GRU hybrid performed better than other models. Despite overfitting on the test set, the Random Forest meta-model demonstrated high accuracy during development.

Despite these encouraging outcomes, issues like overfitting and high processing requirements were found. The study results highlight the need for regularization strategies such as dropout, data augmentation, and cross-validation to mitigate overfitting. For increased accuracy, future studies can investigate transformer-based models like BERT, cross-validation, and sophisticated regularization. Addressing class imbalance through techniques like SMOTE or focal loss will further enhance the robustness of future models.

References

Elisa Bjerre, Michael N Fienen, Raphael Schneider, Julian Koch, and Anker L Højberg. 2022. Assessing spatial transferability of a random forest metamodel for predicting drainage fraction. *Journal of Hydrology*, 612:128177.

Mondher Bouazizi and Tomoaki Ohtsuki. 2016. Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in twitter. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE.

Mondher Bouazizi and Tomoaki Ohtsuki. 2017. A pattern-based approach for multi-class sentiment analysis in twitter. *IEEE Access*, 5:20617–20639.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Ming Jiang, Junlei Wu, Xiangrong Shi, and Min Zhang. 2019. Transformer based memory network for sentiment analysis of web comments. *IEEE Access*, 7:179942–179953.

Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang, and Xiaotie Deng. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

JESSLYN MARCELLINA. 2022. *Metode long short-term memory (LSTM), Gated recurrent unit (GRU), Dan convolutional long short-term memory (CONV-LSTM) untuk peramalan data runtun waktu (Studi Kasus: Jumlah Kasus Positif Harian COVID-19 di Indonesia)*. Ph.D. thesis, Universitas Gadjah Mada.

Dheeraj Mekala, Xinyang Zhang, and Jingbo Shang. 2020. Meta: Metadata-empowered weak supervision for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Erkut Memiş, Hilal Akarkamçı, Mustafa Yeniad, Javad Rahebi, and Jose Manuel Lopez-Guede. 2024. Comparative study for sentiment analysis of financial tweets with deep learning methods. *Applied Sciences*, 14(2):588.

YPPAF Pavel and Brazdil2 Carlos Soares. 2002. Decision tree-based data characterization for meta-learning. *IDDM-2002*, 111.

A Sherin, I Jasmine SelvakumariJeya, and SN Deepa. 2024. Enhanced aquila optimizer combined ensemble bi-lstm-gru with fuzzy emotion extractor for tweet sentiment analysis and classification. *IEEE Access*.

Altyeb Altaher Taha and Sharaf Jameel Malebary. 2022. A hybrid meta-classifier of fuzzy clustering and logistic regression for diabetes prediction. *Computers, Materials & Continua*, 71(3).

Kian Long Tan, Chin Poo Lee, Kalaiarasi Sonai Muthu Anbananthen, and Kian Ming Lim. 2022. Roberta-lstm: a hybrid model for sentiment analysis with

transformer and recurrent neural network. *IEEE Access*, 10:21517–21525.

Lizhen Wu, Chun Kong, Xiaohong Hao, and Wei Chen. 2020. A short-term load forecasting method based on gru-cnn hybrid neural network model. *Mathematical problems in engineering*, 2020(1):1428104.

Alec Yenter and Abhishek Verma. 2017. Deep cnn-lstm with combined kernels from multiple branches for imdb review sentiment analysis. In *2017 IEEE 8th annual ubiquitous computing, electronics and mobile communication conference (UEMCON)*, pages 540–546. IEEE.

Misogynistic Meme Detection in Dravidian Languages Using Kolmogorov Arnold-based Networks

Manasha Arunachalam[1], Navneet Krishna Chukka[1], Harish Vijay V[1]
Premjith B[1], Bharathi Raja Chakravarthi[2]
[1]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India,
[2]School of Computer Science, University of Galway, Ireland,
manasha.arun@gmail.com, navneetkrishna.918@gmail.com
harishvijay0204@gmail.com, b_premjith@cb.amrita.edu
bharathi.raja@universityofgalway.ie

Abstract

The prevalence of misogynistic content online poses significant challenges to ensuring a safe and inclusive digital space for women. This study presents a pipeline to classify online memes as misogynistic or non misogynistic. The pipeline combines contextual image embeddings generated using the Vision Transformer Encoder (ViTE) model with text embeddings extracted from the memes using ModernBERT. These multimodal embeddings were fused and trained using three advanced types of Kolmogorov Artificial Networks (KAN): PyKAN, FastKAN, and Chebyshev KAN. The models were evaluated based on their F1 scores, demonstrating their effectiveness in addressing this issue. This research marks an important step towards reducing offensive online content, promoting safer and more respectful interactions in the digital world.

1 Introduction

Recent studies have highlighted the role of social media algorithms in amplifying such harmful content, thereby normalizing detrimental ideologies among users. Addressing this issue necessitates effective detection and mitigation strategies. In this study, we propose a comprehensive pipeline for classifying online memes as either containing misogynistic content or not. This approach integrates multimodal data by combining contextual image embeddings from the Vision Transformer Encoder (ViTE) model with text embeddings derived from ModernBERT. The fused embeddings are then processed through advanced Kolmogorov-Arnold Networks (KAN), specifically PyKAN, FastKAN, and Chebyshev KAN, to enhance classification accuracy. The efficacy of these models is evaluated using F1 scores, demonstrating their potential in identifying and mitigating offensive online content. This research helps create safer and more inclusive digital environments by providing a robust method for detecting misogynistic material.

Kolmogorov-Arnold Network (KAN): A modern neural network architecture built on the Kolmogorov-Arnold representation theorem, which asserts that any continuous multivariate function can be decomposed into a combination of single-variable functions. Using a two-layer structure, KAN approximates the target function by combining input mappings to a higher-dimensional space in the inner layer. In KAN, the activation function consists of a spline function, parameterized as a linear combination of B-splines, a base function, often referred to as the SiLU (Sigmoid Linear Unit). KAN is theoretically robust, but because of the complexity of the spline function, it can be computationally difficult to train.

Chebyshev KAN :The Kolmogorov-Arnold theorem is expanded upon by Chebyshev KAN, which uses Chebyshev polynomials for function approximation. Using a single-layer Chebyshev interpolation method, KAN models the target function through a weighted sum of Chebyshev polynomials, with the input normalized by a hyperbolic tangent (tanh) function. While following the theoretical underpinnings of the Kolmogorov-Arnold theorem, ChebyKAN takes advantage of the superior approximation properties of Chebyshev polynomials. This method seeks to improve function approximation's accuracy and efficiency in comparison to the original KAN.

FastKAN: In order to overcome the computational difficulties of KAN, FastKAN substitutes B-spline basis with Gaussian Radial Basis Functions (RBFs), greatly lowering the computational cost. B-splines can be efficiently approximated by Gaussian RBFs, which removes the computational bottleneck in the original KAN implementation. Because of this modification, FastKAN is more feasible for real-time and large-scale applications while preserving the approximation capabilities of KAN and increasing computing efficiency.

2 Literature Review

Previous research (Ponnusamy et al., 2024) has focused on detecting misogyny and gender bias in online spaces, with datasets in languages like English and Spanish. However, there is a lack of studies focusing on regional languages, particularly Tamil and Malayalam, where cultural context plays a significant role. Existing tools often don't address these specific issues effectively. The MDMD dataset fills this gap by providing a resource for detecting misogyny in Tamil and Malayalam memes, helping researchers understand and address gender bias in these communities more accurately.

In recent years, detecting toxic and abusive comments on social media has become crucial to maintaining a safe online environment. Several models have been developed to identify hate speech, toxicity, and bullying, primarily in high-resource languages like English. However, there is limited research on detecting such content in low-resource languages, such as Tamil. Previous work has highlighted the challenges of language-specific nuances, especially when it comes to understanding cultural contexts. This paper (Bhattacharyya, 2022) contributes to the gap by focusing on Tamil, approaching the problem of abusive comment detection as a multi-class classification task. The study compares various pre-processing and modeling techniques, evaluating their effectiveness based on weighted average accuracy.

Recent research (Shaun et al., 2024) has explored the classification of Tamil and Malayalam memes as misogynistic or non-misogynistic. One approach involved separately analyzing textual content using Multinomial Naive Bayes and visual content using the ResNet50 model. By combining the results from both modalities, researchers achieved significant success in identifying misogynistic content in memes. This work underscores the importance of multi-modal analysis in detecting harmful content, especially in low-resource languages.

Detecting misogynistic memes is challenging due to the complex interaction between image and text, where these elements often convey different meanings.Prior research (Jindal et al., 2024) has focused on individual modalities, such as text or image analysis, but these approaches overlook the need for multimodal fusion. Recent works have started exploring fusion techniques, utilizing models like Vision Transformer for images and transformer-based models like DistilBERT for text. These approaches have shown promise in improving the detection of harmful content. However, there remains a gap in combining these modalities effectively, especially in detecting misogyny. The MISTRA framework addresses this by using variational autoencoders for dimensionality reduction and large language models for fusion embeddings, enhancing classification performance on multimodal data.

Additional Studies (Sharma et al., 2024) have explored the use of deep learning models, such as recurrent neural networks (RNN), long-short term memory (LSTM), and bidirectional LSTM, for detecting various categories of hate speech, including misogyny, misandry, and xenophobia. These models are applied to Tamil and Tamil-English code-mixed comments, and results are analyzed to evaluate their effectiveness in identifying abusive content.

Social media memes, combining text and images, can sometimes contain harmful content like misogyny, affecting users' well-being. Detecting such content, especially in low-resource languages, is challenging due to the lack of suitable datasets. This work (Singh et al., 2024) introduces a Hindi-English code mixed meme dataset of 5,054 annotated memes for two tasks: misogyny detection and multi-label classification. Results show that multimodal fusion models outperform text-only and image-only models in identifying misogyny. This dataset provides a valuable resource for advancing research in detecting harmful online content.

Misogynistic memes, which target women with disrespectful language, pose a challenge to maintaining a healthy online environment. The paper (Mahesh et al., 2024) presents three models: BERT+ResNet-50, MuRIL+ResNet-50, and mBERT+ResNet-50, which combine text and image representations for meme classification. The mBERT+ResNet-50 and MuRIL+ResNet-50 models achieved impressive macro F1 scores of 0.73 and 0.87 for Tamil and Malayalam datasets, securing 1st place for both languages in the shared task.

A system for identifying abusive remarks in Tamil and Tamil-English is shown in the work (Duraphe et al., 2022) utilizing three different approaches: transformer-based modeling, deep learning, and machine learning. Classifying remarks into groups such as misogyny, misandry, homophobia, and others was their goal. For the Tamil+English dataset, the system performs best

when employing Random Forest, with a weighted average F1-score of 0.78. Furthermore, mBERT produces the best result for Tamil with an F1-score of 0.7 in Transformer-based modeling, whereas Bi-Directional LSTM performs better for Deep Learning.

A unique method for identifying inappropriate language on social media in multilingual, code-mixed, and script-mixed contexts is presented in this study (Saumya et al., 2024). The challenge makes use of a hybrid multilingual dataset that was produced by fusing bilingual and monolingual materials. The study assesses the effects of deep learning models (CNN, Bi-LSTM, Bi-LSTM-Attention, and fine-tuned BERT) and various input representations (Word2Vec, GloVe, BERT, and uniform initialization). With a macro average F1-score of 0.79 for monolingual tasks and 0.86 for code-mixed/script-mixed tasks, the results show how well fine-tuned BERT performs, improving the identification of abusive language in a variety of multilingual contexts.

3 Dataset Description

The Misogyny Meme Detection dataset consists of images, corresponding transcriptions, and labels (Ponnusamy et al., 2024).

For Tamil, the training data includes 1,135 images and a CSV file containing image IDs, transcriptions, and labels (0 or 1). Out of these, 732 images have matching entries in the CSV file based on image IDs. The development (dev) set contains approximately 282 images and a similar CSV file with image IDs, transcriptions, and labels . Among these, 252 images have matching entries in the CSV file based on image IDs. The test set comprises approximately 356 images, along with a CSV file containing transcriptions and labels. Among these, 82 images have corresponding entries in the CSV file based on their image IDs. The images and their corresponding transcriptions were used to predict the labels, and the predicted labels were evaluated against the true labels provided in the CSV file to measure accuracy.

In the Tamil dataset, some images listed in the test set were missing from the provided test data. To address this, the missing images were searched for in the training and development sets. Once identified, these images were added to the test set. It was confirmed that these images had not been used during training or validation, ensuring the test set remained unique and independent. A similar approach was applied to the training and development sets. For the training set, missing images were identified by cross-checking the data entries and were searched for in the development set. Likewise, for the development set, any missing images were located in the training set. This ensured that all data was appropriately assigned while maintaining the uniqueness and integrity of each dataset.

The Malayalam Misogyny Meme Detection dataset includes 640 images in the training set, accompanied by a CSV file containing image IDs, transcriptions, and labels (0 or 1). The development (dev) set consists of 160 images and a corresponding CSV file with the same structure. The test set contains 200 images and a CSV file with transcriptions and labels. Similar to Tamil, the images and transcriptions in the test set were used to predict the labels, and the accuracy of these predictions was evaluated by comparing them with the true labels from the CSV file.

4 Methodology

As per Figure(1), For the Tamil and Malayalam dataset, we used training data that contained images and their corresponding transcriptions, along with binary labels (0 or 1).

For each image in the dataset, we used the Vision Transformer Encoder (ViTE) model to extract high-dimensional embeddings. This allowed us to represent visual data effectively for downstream tasks.

The transcriptions associated with each image were processed using ModernBERT, which generated text embeddings representing the semantic content of the text. The image and text embeddings were fused to create a single representation for each data point. This fused embedding served as the input for the classification models. We trained three types of Kolmogorov Artificial Networks (KANs) using the fused embeddings: PyKAN, Chebyshev KAN and FastKAN and fine tuned using the hyperparameters listed in Table 1.

We utilized ModernBERT to generate text embeddings. ModernBERT is an advanced language model that enhances the original BERT architecture by extending the context length to 8,192 tokens, allowing it to process longer documents effectively. It incorporates Rotary Positional Embeddings (RoPE) for improved token position understanding and replaces traditional MLP layers

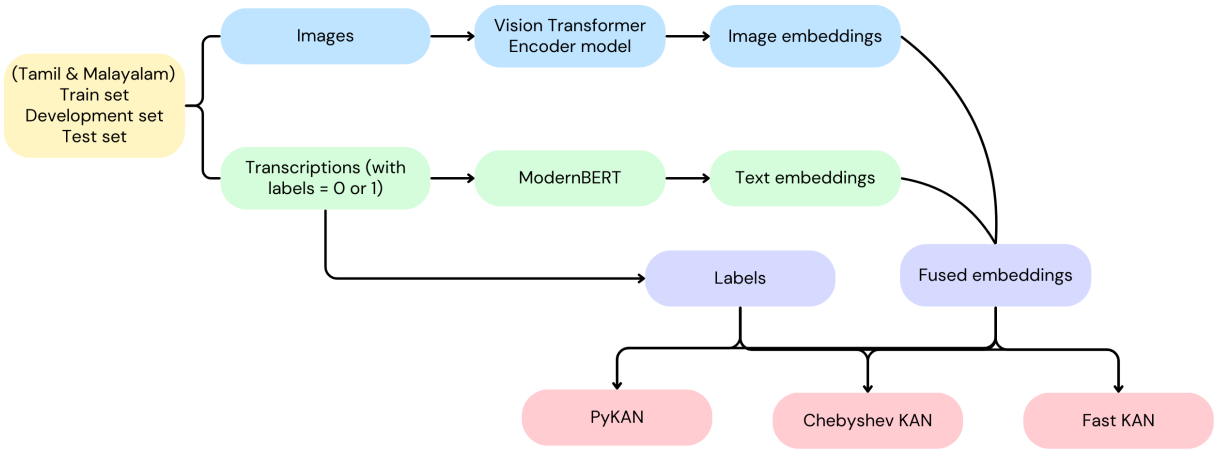


Figure 1: Overall Workflow

with GeGLU layers, enhancing model performance. These architectural improvements allow ModernBERT to produce more comprehensive and contextually rich embeddings, which is crucial for our classification task.

To generate image embeddings, we employed the Vision Transformer Encoder (Remya et al., 2024) (ViTE) model. ViTE operates by dividing an input image into fixed-size patches, each of which is linearly transformed into a vector representation. These patch embeddings are then combined with positional embeddings to retain spatial information. The resulting sequence is processed through transformer encoder layers, enabling the model to capture both local and global features of the image. This method allows ViTE to generate comprehensive embeddings that effectively represent the visual content.

5 Experiments and Discussion

5.1 Experimental Setup

The experiments were carried out on a MacBook (M4 Pro) equipped with 24GB of Unified Memory and a 512GB SSD. This setup, combined with PyTorch's support for macOS using the Metal Performance Shaders (MPS) backend, allowed for smooth model loading and faster training by efficiently utilizing the Mac hardware capabilities.

5.2 Hyperparameters

The hyperparameters used for training the KAN models have been discussed below in table 1.

Model	Learning Rate	Epochs	Batch Size	Optimiser
Chabyshev KAN (Tamil)	0.0001	20	32	Adam
PyKAN (Tamil)	0.001	20	32	Adam
FastKAN (Tamil)	0.001	55	32	Adam
Chabyshev KAN (Malayalam)	0.0001	20	32	Adam
PyKAN (Malayalam)	0.001	20	32	Adam
FastKAN (Malayalam)	0.001	55	32	Adam

Table 1: Hyperparameters and Optimisers for Different Models

5.3 Software packages

We used PyTorch and TensorFlow for model training and related tasks, while scikit-learn was employed for evaluation metrics. PyTorch and TensorFlow are prominent deep learning frameworks that facilitate the development and training of neural networks. Scikit-learn, on the other hand, is a widely-used machine learning library that provides tools for data analysis and model evaluation.

6 Results

The F1 Score—the harmonic mean of precision and recall—was the main metric we used to assess our models. Because it ensures that both false positives and false negatives are taken into account, it is especially helpful in situations where the dataset is unbalanced. In our case, we found that class 1, the misogynistic class, had less incidents. The following formula provides the F1 score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

The ratio of accurately predicted positive observations to all actual positives is called recall, while the ratio of properly predicted positive observations to all predicted positives is called precision.

The F1 Score is preferred in our evaluation for the following reasons: It balances precision and recall, making it suitable for imbalanced datasets where accuracy alone may be misleading. It ensures that both false positives and false negatives are accounted for, which is crucial for our classification task. Unlike accuracy, F1 Score does not get skewed when one class is significantly larger than the other. Thus, using the F1 score provides a more reliable measure of the model's effectiveness in real-world scenarios like these.

6.1 Results for Tamil

As observed in Table 1,2,3, the highest F1-score of 0.77 for ChebysevKAN, followed by 0.76 for PyKAN and 0.73 for FastKAN.

ChebysevKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.88	0.90	0.89	267
1.0	0.67	0.63	0.65	89
Accuracy		0.83		356
Macro Avg	0.77	0.76	**0.77**	356
Weighted Avg	0.83	0.83	0.83	356

Table 2: Test Classification Report for ChebyshevKAN on Tamil dataset

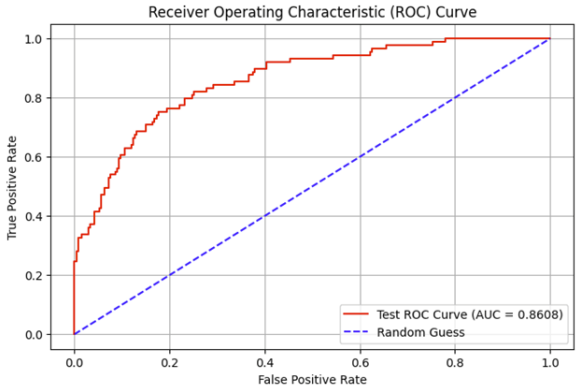


Figure 2: ROC curve for ChebysevKAN on Tamil dataset

PyKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.88	0.87	0.88	267
1.0	0.63	0.65	0.64	89
Accuracy		0.82		356
Macro Avg	0.76	0.76	**0.76**	356
Weighted Avg	0.82	0.82	0.82	356

Table 3: Test Classification Report for PyKAN on Tamil dataset

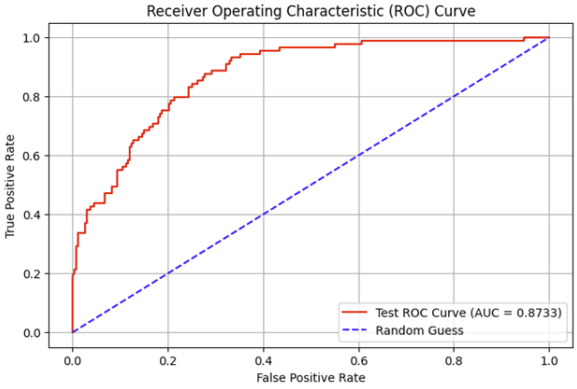


Figure 3: ROC cureve for PyKAN on Tamil dataset

FastKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.84	0.94	0.89	267
1.0	0.74	0.47	0.58	89
Accuracy		0.83		356
Macro Avg	0.79	0.71	**0.73**	356
Weighted Avg	0.82	0.83	0.81	356

Table 4: Test Classification Report for FastKAN on Tamil dataset

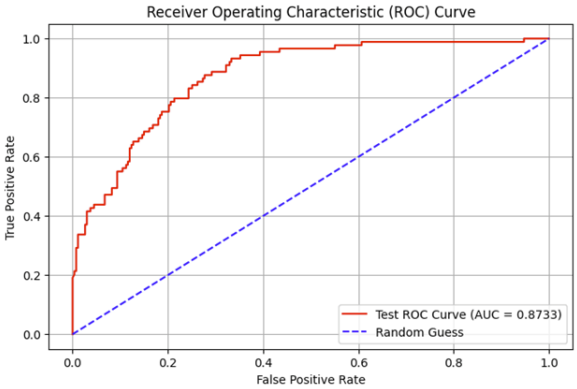


Figure 4: ROC curve for FastKAN on Tamil dataset

All three models achieve similar accuracy (around 0.83) and perform well on class 0, with high precision and recall (above 0.84). Each model demonstrates strengths and weaknesses depending on the classification task.

As observed from Figure (2,3,4) Chebyshev KAN offers the best balance, PyKAN is a close alternative, and Fast KAN performs well but needs recall improvements for class 1. KAN models misclassifies a higher number of class 1 samples, likely due to feature overlaps or data imbalance. Further tuning and data adjustments can enhance overall performance, particularly for class 1 detection.

6.2 Results for Malayalam

The same model architecture was applied on Malayalam dataset. The results for each model (FastKAN, ChebysevKAN, and PyKAN) are presented below, along with their respective classification reports and ROC curves.

All three models (FastKAN, ChebysevKAN, and PyKAN) from Table 5,6,7, The models demonstrated strong performance on the Malayalam dataset, with FastKAN achieving the highest accuracy (0.88) and F1-score (0.87). ChebysevKAN and PyKAN followed closely, with accuracies of 0.87 and 0.86, respectively. The ROC curves for all models as observed from Figure (5,6,7) indicate excellent discrimination capabilities, with high AUC values.

Additionally, the classification reports reveal that FastKAN consistently achieved higher recall for class 1, making it particularly effective in identifying positive instances. ChebyshevKAN demonstrated a more balanced performance across both classes, while PyKAN exhibited slightly lower recall but maintained competitive precision. The ROC curves further validate the model's classification capabilities, with all AUC values exceeding 0.85.

These findings suggest that the proposed architecture not only generalizes well across different languages but also adapts effectively to the nuances of the Malayalam dataset, with FastKAN being the most effective among the three.

FastKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.88	0.93	0.90	122
1.0	0.87	0.79	0.83	78
Accuracy		0.88		200
Macro Avg	0.87	0.86	**0.87**	200
Weighted Avg	0.87	0.88	0.87	200

Table 5: Test Classification Report for FastKAN on Malayalam dataset

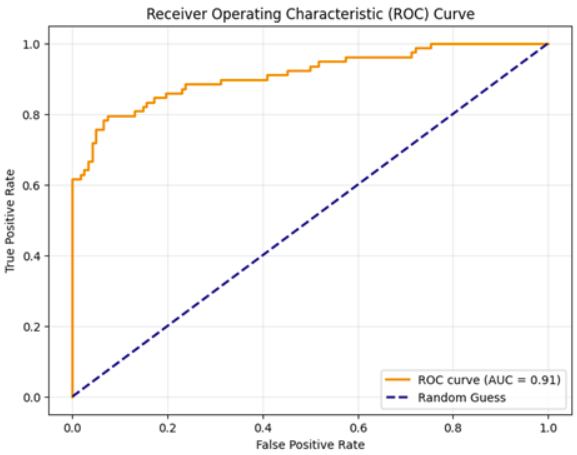


Figure 5: ROC curve for FastKAN on Malayalam data

ChebysevKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.87	0.93	0.90	122
1.0	0.87	0.78	0.82	78
Accuracy		0.87		200
Macro Avg	0.87	0.85	**0.86**	200
Weighted Avg	0.87	0.87	0.87	200

Table 6: Test Classification Report for ChebysevKAN on Malayalam dataset

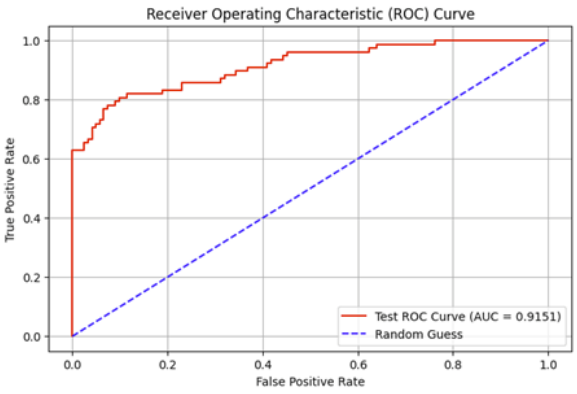


Figure 6: ROC curve for ChebysevKAN on Malayalam dataset

PyKAN

Class	Precision	Recall	F1-Score	Support
0.0	0.88	0.89	0.89	122
1.0	0.83	0.81	0.82	78
Accuracy		0.86		200
Macro Avg	0.85	0.85	**0.85**	200
Weighted Avg	0.86	0.86	0.86	200

Table 7: Test Classification Report for PyKAN on Malayalam dataset

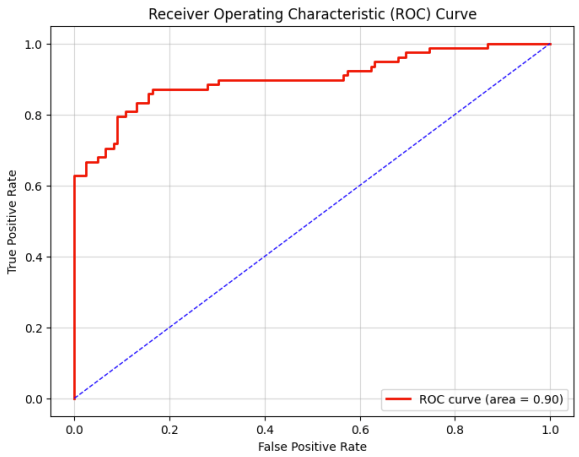


Figure 7: ROC curve for PyKAN on Malayalam dataset

To mitigate the class imbalance in our dataset, where misogynistic instances were significantly underrepresented compared to non-misogynistic ones, we used various oversampling techniques to generate a more balanced distribution. Specifically, we applied SMOTE (Synthetic Minority Oversampling Technique), KMeansSMOTE, ADASYN (Adaptive Synthetic Sampling), and BorderlineSMOTE, each of which synthesizes new samples for the minority class using different interpolation strategies. These methods allowed us to expand the misogynistic class while preserving the overall data distribution, ensuring that our KAN models had sufficient representative samples to learn nuanced patterns associated with misogynistic content.

After incorporating the oversampled data into our training pipeline, we retrained our KAN models and observed that the overall performance remained comparable to our initial results. However, the class balance improved significantly, leading to an increase in key metrics for the misogynistic class (Class 1). This indicates that the model's ability to correctly identify misogynistic content improved without introducing substantial biases toward the majority class. The results highlight the effectiveness of oversampling in handling class imbalances and ensuring better model performance across both classes.

7 Inference

From the classification reports, we observe that all models achieve high precision and recall for class 0 (non-misogynistic text) but exhibit lower recall for class 1 (misogynistic text), indicating that they struggle to correctly identify some misogynistic instances.

Among the models, ChebyshevKAN performed best relative to F1 score on Tamil dataset and FastKAN on Malayalam dataset.

We could enhance recall through advanced feature engineering, leveraging larger and more diverse datasets, or incorporating multimodal approaches to improve robustness and fairness in misogyny detection.

8 Conclusion

This study shows that combining image and text features using Vision Transformer Encoder (ViTE) and ModernBERT with Kolmogorov Arnold Networks (KAN) is effective for detecting misogynistic memes in Tamil and Malayalam. Among the models tested, FastKAN performed best for Malayalam (F1 score: 87), while Chebyshev KAN was the most effective for Tamil (F1 score: 77). Despite the class imbalance—where misogynistic content was underrepresented in the dataset—the models achieved reasonable scores, demonstrating their robustness. However, further fine-tuning is needed to address the challenges posed by this imbalance.

Future work can focus on expanding the dataset to more Indian languages, improving fusion techniques for better accuracy, and optimizing models for real-time deployment in social media moderation tools. Additionally, incorporating techniques to handle class imbalance and integrating bias detection and explainability methods can make the system more transparent and fair. These advancements will help in automating content moderation, reducing harmful content, and promoting inclusive online communities.

References

Aanisha Bhattacharyya. 2022. Aanisha@ tamilnlp-acl2022: abusive detection in tamil. In *Proceedings*

of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 214–220.

Ankita Duraphe, Ratnavel Rajalakshmi, and Antonette Shibani. 2022. Dlrg@ dravidianlangtech-acl2022: Abusive comment detection in tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics (ACL).

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Sidharth Mahesh, D Sonith, Gauthamraj Gauthamraj, G Kavya, Asha Hegde, and H Shashirekha. 2024. Mucs@ lt-edi-2024: Exploring joint representation for memes classification. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 282–287.

Rahul Ponnusamy, Kathiravan Pannerselvam, R Saranya, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, S Bhuvaneswari, Anshid Ka, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in tamil and malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488.

S Remya, T Anjali, S Abhishek, Somula Ramasubbareddy, and Yongyun Cho. 2024. The power of vision transformers and acoustic sensors for cotton pest detection. *IEEE Open Journal of the Computer Society*.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2024. Filtering offensive language from multilingual social media contents: A deep learning approach. *Engineering Applications of Artificial Intelligence*, 133:108159.

Deepawali Sharma, Vedika Gupta, and Vivek Kumar Singh. 2024. Abusive comment detection in tamil using deep learning. In *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications*, pages 207–226. Elsevier.

H Shaun, Samyuktaa Sivakumar, R Rohan, Nikilesh Jayaguptha, and Durairaj Thenmozhi. 2024. Quartet@ lt-edi 2024: A svm-resnet50 approach for multitask meme classification-unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 221–226.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Detection of Religious Hate Speech During Elections in Karnataka

MSVPJ Sathvik
Raickers AI
Hyderabad
Telangana, India
msvpjsathvik@gmail.com

Raj Sonani
Cornell University
Ithaca
New York, USA
sonaniraj@gmail.com

Ravi Teja Potla
Slalom
Houston
Texas, USA
raviteja.potla@gmail.com

Abstract

We propose a novel dataset for detecting religious hate speech in the context of elections in Karnataka, with a particular focus on Kannada and Kannada-English code-mixed text. The data was collected during the Karnataka state elections and includes 3,000 labeled samples that reflect various forms of online discourse related to religion. This dataset aims to address the growing concern of religious intolerance and hate speech during election periods, it's a dataset of multilingual, code-mixed language. To evaluate the effectiveness of this dataset, we benchmarked it using the latest state-of-the-art algorithms. We achieved accuracy of 78.61%.

1 Introduction

Religious tensions between Hindus and Muslims have been a sensitive issue in India, often increasing during elections(Pradhan and Mehta, 2019). In Karnataka, some political parties have been accused of spreading religious hatred to gain votes. This has led to violent incidents, communal clashes, and even loss of lives. Social media plays a major role in amplifying hate speech, as people use these platforms to express strong opinions, sometimes leading to misinformation, targeted attacks, and communal propaganda(Kumar and Gupta, 2020).

During elections, the amount of Hindu-Muslim hate speech on social media rises sharply(Narayanan et al., 2019). Many users post content that provokes religious sentiments, causing division and unrest. Despite social media companies trying to control harmful content, their existing detection systems struggle with regional languages and code-mixed text. Kannada and Kannada-English code-mixed speech make it even harder for AI models to identify hate speech accurately.

Motivation: Political campaigns often intensify religious, ethnic, and ideological divisions, leading to social unrest and real-world violence. Elections are a time when public opinion is highly influenced, and the spread of hate speech on social media can manipulate voters, incite communal tensions, and weaken democratic values. Unchecked hate speech can lead to misinformation, voter suppression, targeted harassment, and even violent clashes between communities. In regions like Karnataka, where religious polarization is sometimes exploited for political gains, identifying and controlling hate speech can prevent riots, protect vulnerable communities, and ensure fair and peaceful elections.

How can we detect hate speech? that too for state elections of Karnataka? Kannada is a low-resource language, and social media conversations often involve Kannada-English code-mixed text. While AI can help predict hate speech, the absence of a reliable dataset makes it difficult to develop and evaluate effective models. Although advanced AI models exist, there is no clear understanding of which model performs best for this specific task. To address this gap, we propose a novel dataset tailored for Hindu-Muslim hate speech detection in Kannada and Kannada-English text. Additionally, we benchmark this dataset using state-of-the-art large language models (LLMs) to compare their effectiveness, providing valuable insights into the most suitable AI model for detecting hate speech in regional and code-mixed languages.

Our key contributions are as follows:

1. As of our knowledge we are the first to develop a dataset for the detection of the hate speech on religious issues during elections.

2. We have benchmarked the dataset with SOTA models and presented the results and comparison.

2 Related Work

There are several research works focused on Kannada hate speech detection. Chakravarthi et al. (2021) introduced the Dravidian CodeMix dataset,

which includes Kannada-English code-mixed text, allowing researchers to develop language models capable of handling mixed-script data. However, the dataset primarily focuses on sentiment classification rather than explicit hate speech detection. Patil et al. (2022) created a Kannada hate speech dataset from social media posts, demonstrating that transformer-based models like mBERT outperform traditional models like SVM and LSTMs in this task.Suryawanshi et al. (2020) built a dataset for Tamil-English code-mixed sentiment analysis, which provided valuable insights into handling mixed-language text. Extending such techniques to Kannada-English code-mixed data is crucial for improving detection models. Ramesh et al. (2023) proposed a hybrid deep learning model combining LSTMs with attention mechanisms for detecting hate speech in Tamil-English and Telugu-English code-mixed tweets. Their findings indicate that context-aware embeddings such as IndicBERT significantly improve performance,

Risch et al. (2021) explored offensive language detection using multilingual transformer models, concluding that fine-tuning models on code-mixed and regional datasets significantly enhances performance. This aligns with recent efforts to apply pre-trained multilingual models like XLM-R and IndicBERT for Kannada hate speech detection.

3 Methodology

3.1 Data Collection and Annotation

For this study, we collected data from Twitter using the Twitter API, ensuring that the dataset includes real-time social media conversations in both Kannada and Kannada-English code-mixed text. The tweets were filtered based on keywords, hashtags, and engagement metrics to capture a diverse set of opinions and discussions related to Hindu-Muslim hate speech.

To ensure high-quality annotations, we employed a team of three native Kannada-speaking annotators who were responsible for labeling the dataset. Each annotator was provided with an Excel sheet containing the collected data, and they were instructed to label each text instance as either hate speech on religion (1) or non-hate speech (0). To maintain annotation consistency and reliability, each data point was labeled by at least two annotators, ensuring that disagreements could be reviewed and resolved.

To measure the Inter-Annotator Agreement (IAA), we calculated pairwise agreement scores between annotators. The agreement scores were as follows: I(1,2) = 87.2%, I(2,3) = 89.6%, and I(1,3) = 86.1%, demonstrating strong agreement among the annotators. These high agreement values indicate that the dataset is well-annotated and reliable, making it suitable for training and evaluating AI models for religious hate speech detection.

Table 1: Statistics of the Dataset

Metrics	label 0	label 1	Total/Overall
Data Size	1542	1633	3175
Number of Words	38103	42294	80397
Words per data point	24.71	25.90	25.32

Data Annotation Guidelines

To ensure high-quality labeling, three native Kannada-speaking annotators followed strict annotation guidelines. Each text sample was labeled based on the presence of religious hate speech. The annotators were given with the data points and were given to label each and every data point accordingly.

Labeling Criteria:

- **Hate Speech on Religion (1)**: Texts containing explicit or implicit hate against a religion (Hindu or Muslim), including:
  - Direct insults, slurs, or abusive language targeting a religious group.
  - Calls for violence, discrimination, or exclusion based on religion.
  - Misinformation or stereotypes that incite hatred.
  - Provocative statements intending to cause communal disharmony.
- **Not Hate Speech (0)**: Texts that do not contain religious hate speech, including:
  - Political discussions without religious targeting.
  - Opinions, sarcasm, or humor that do not promote hatred.
  - Discussions about religion without intent to incite violence.

If a sample was unclear, annotators referred to context and intent. Disagreements between two

Table 2: Overview of the dataset

Text	Label[0/1]
Deshada Hindu matthu Muslim galu yendu ondu ge barolla avaru olage namma rashtrada vairi Muslim galu namma desh dushman	1
Muslim galu hamesha Hindu mele droh madtare ivaru namma desh ke khatara ivaranna olage vittu namma rashtrada surakshya dodda aapathige nuggi bidutte	1
Kannada naadu belongs to all religions we must promote peace and harmony society flourishes when we stand together respecting all faiths and backgrounds equally	0
Muslim galu matthu Hindu galu seridre matra naadu balutte, yava ondu dharmavannu dushisuvudu sari alla, namma sahane namma balavagi iruvudu.	0
Religious harmony is important for a strong nation, Hindus and Muslims must coexist peacefully, respecting each other's traditions, ensuring equality, and spreading love, not hate.	0
Muslims get out of our country this land belongs to Hindus they are a threat to our nation they should leave immediately we cannot trust them anymore	1

annotators were resolved through discussions. Misinformation leading to potential hate was labeled as hate speech.

3.2 Analysis

Table 1 presents the statistical analysis of the dataset, 3,175 data points with 1,542 labeled as non-hate speech (0) and 1,633 as hate speech (1), ensuring a balanced distribution for training AI models. The dataset contains 80,397 words, with hate speech samples contributing 42,294 words and non-hate speech 38,103 words (47.4%), hence the dataset is balanced. Table 2 represents the overview of the dataset which displays few examples from the proposed dataset.

3.3 Baselines

We conducted experiments on the proposed dataset using various pre-trained language models and large language models (LLMs), including: (i) GPT-4o(OpenAI, 2023), (ii) Gemini(DeepMind, 2023), (iii) LLaMA 3(Touvron et al., 2023), (iv) Kannada-BERT(Khanuja et al., 2021), (v) IndicBERT(Kakwani et al., 2020), and (vi) Multilingual-BERT(Devlin et al., 2018).

For baseline experimentation, we implemented the few-shot prompting technique, where eight training samples were selected from the dataset. These examples were provided as context to guide the LLMs in classifying the input text.

The dataset was randomly split into 80% for training and 20% for testing. Pre-trained models were fine-tuned for five epochs with a learning rate of 0.01, while other parameters were kept at default settings. GPT variants were fine-tuned using the OpenAI API key, while BERT-based models were fine-tuned on Google Colab (free GPU version). Few-shot prompting was executed without GPU on Google Colab, whereas LLaMA models were fine-tuned using NVIDIA GPUs with CUDA support.

Table 3: Test results

Model	Precision	Recall	Accuracy
LSTM	58.42	60.19	59.32
Bi-LSTM	62.71	67.02	65.14
CNN Bi LSTM	64.12	70.15	68.22
m-BERT	68.35	72.14	71.48
Kannada BERT	70.92	75.34	73.59
Indic BERT	72.54	76.92	75.36
Gemini 2.0	75.18	79.11	77.51
LLAMA 3	76.24	77.61	77.02
GPT-4o	78.36	79.81	78.61

4 Experimental Results and Discussion

Table 3 presents the test results of several models across precision, recall, and accuracy. In general, older models like LSTM, Bi-LSTM, and CNN Bi LSTM have lower accuracy compared to the more advanced models. Among these, CNN Bi LSTM achieves the highest accuracy but still falls short of the more recent models.

When it comes to transformer-based models, m-BERT shows a noticeable improvement over the earlier models. Its accuracy is higher than the LSTM-based models, indicating that transformer architectures tend to perform better for the given task. Kannada BERT and Indic BERT further outperform m-BERT, with Indic BERT reaching the highest accuracy among the BERT-based models.

The highest accuracy scores are seen in the latest models Gemini 2.0, LLAMA 3, and GPT-4o.

These models significantly surpass the accuracy of the previous ones, with GPT-4o achieving the highest accuracy overall. This suggests that the most recent developments in transformer-based models, particularly those like GPT-4o, are highly effective for the task and represent a significant leap in performance.

Real time usecases:

The models trained on this dataset can be applied to various real-time scenarios to help manage and control hate speech online, particularly during sensitive times like elections.

1. **Election Commission:** The model can assist the Election Commission in identifying and removing posts that spread religious hate during election periods. This helps maintain a peaceful and unbiased environment, ensuring that elections remain fair and free from divisive content. For example, if a social media post targets a particular religious group with inflammatory remarks, the system can flag it for removal, promoting a more respectful and impartial electoral process.

2. **Social Media Platforms:** Social media companies can use this model to monitor and regulate content that may negatively influence teenagers and children. As young people are more susceptible to harmful content, the system can help identify and restrict posts that spread religious intolerance or hate speech. For instance, if a user posts a hate-filled comment targeting a minority religious group, the platform could use the model to detect it and either warn the user or remove the post to protect younger audiences.

3. **Government Agencies and Law Enforcement:** The model could be used by government bodies or law enforcement agencies to track and prevent the spread of hate speech across public forums, particularly during sensitive times like political unrest or elections. By detecting harmful content early, agencies can take proactive measures to prevent violent outbreaks or social division. For example, it could help identify extremist posts before they escalate into offline actions.

4. **Media and News Outlets:** News organizations could use this model to monitor and manage the spread of biased or harmful religious narratives in the media, especially during election seasons when the risk of divisive rhetoric is higher. By detecting inflammatory language, media outlets can avoid amplifying hate speech or biased content in their reports, ensuring a more balanced and responsible approach to news coverage.

5. **Educational Institutions:** Schools and universities can use the model to monitor online discussions, forums, or social media groups where students interact. This helps maintain a safe and inclusive environment by identifying and addressing harmful content related to religion, fostering respectful discourse among young people. For example, a student might post hateful comments about another religion in an online forum, and the system can flag it for review by administrators.

5 Conclusion and Future Work

The proposed hate speech detection model has significant potential to mitigate religious hate speech in real-time, particularly during elections and on social media platforms. By using advanced natural language processing (NLP) techniques, this model helps various stakeholders, including election commissions, social media companies, law enforcement agencies, and educational institutions, to identify and control harmful speech. The real-world applications discussed demonstrate the necessity of such models in maintaining a fair, unbiased, and safer online environment. Future work includes developing advanced algorithms for detecting hate speech in Kannada and other Dravidian languages while also expanding the scope to explore related issues such as caste politics and other forms of social discrimination.

Limitations

This study primarily focuses on text-based hate speech detection, which means it does not account for other modalities such as images, videos, or audio, where hate speech can also be prevalent. Additionally, the scope of this research is centered around elections, providing valuable insights into political discourse but not extending to other important social contexts such as caste-based discrimination or general communal hate speech outside election periods. However, this focused approach

allows for a deeper and more precise understanding of election-related hate speech, laying the groundwork for future research to expand into multimodal analysis and broader societal issues.

Ethics Statment

We strongly oppose any potential misuse of this dataset, such as training models to generate hate speech or promote religious discrimination. Our sole aim is to detect religious hate speech during elections and help mitigate its spread on social media, fostering a safer and more inclusive online environment.

References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Jishnu Jose, Thomas Mandl, Mitesh M. Kumar, and Elizabeth Sherly. 2021. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. In *Proceedings of the 13th International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 714–722.

Google DeepMind. 2023. Gemini 2.0: Scaling and advances in large language models. *Google AI Blog*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhruva Kakwani, Anoop Kunchukuttan, Shiva Meena Golla, Pushpak Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Simran Khanuja, Sandipan Dandapat, Ritesh Kumar, Sunayana Sitaram, K. P. Soman, and Anup Kumar. 2021. Mahanlp: Towards indic language understanding using bert models for hindi, marathi, and kannada. *arXiv preprint arXiv:2106.07469*.

Rajesh Kumar and Anjali Gupta. 2020. The role of social media in spreading religious hate speech during elections in india. *International Journal of Communication and Society*, 5(2):220–235.

Vidya Narayanan, Vladimir Barash, Bence Kollanyi, Lisa-Maria Neudert, and Philip Howard. 2019. News and information overload in the indian elections: The case of whatsapp. *Computational Propaganda Project, Oxford Internet Institute*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aniket Patil, Raghav Sharma, and Anil Kumar. 2022. Kannada hate speech detection using transformer-based models. *Journal of Computational Linguistics and Artificial Intelligence*, 8(2):102–118.

Pratyush Pradhan and Sanjay Mehta. 2019. Religious polarization and electoral politics in india. *Indian Journal of Political Science*, 80(3):345–362.

S. Ramesh, V. Kumar, P. Srinivasan, and A. Iyer. 2023. Hybrid deep learning model for hate speech detection in tamil-english and telugu-english code-mixed text. *International Journal of Computational Linguistics and NLP*, 10(2):45–62.

Julian Risch, Anke Stoll, and Ralf Krestel. 2021. Offensive language detection exploiting multilingual transformer models. *Natural Language Processing Journal*, 35(4):567–589.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Karun Arora, Elizabeth Sherly, and John P. McCrae. 2020. A dataset for sentiment analysis of code-mixed tamil-english text. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 2459–2468.

Hugo Touvron, Louis Martin, Kevin Stone, Pierre Albert, Amjad Almahairi, Ross Taylor, Gautier Izacard, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

DravLingua@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages using Late Fusion of Muril and Wav2Vec Models

Aishwarya S
Sri Eshwar College of Engineering
Coimbatore, Tamil Nadu, India
aishwarya.s2020cse@sece.ac.in

Abstract

Detecting hate speech on social media is increasingly difficult, particularly in low-resource Dravidian languages such as Tamil, Telugu and Malayalam. Traditional approaches primarily rely on text-based classification, often overlooking the multimodal nature of online communication, where speech plays a pivotal role in spreading hate speech. We propose a multimodal hate speech detection model using a late fusion technique that integrates Wav2Vec 2.0 for speech processing and Muril for text analysis. Our model is evaluated on the DravidianLangTech@NAACL 2025 dataset, which contains speech and text data in Telugu, Tamil, and Malayalam scripts. The dataset is categorized into six classes: Non-Hate, Gender Hate, Political Hate, Religious Hate, Religious Defamation, and Personal Defamation. To address class imbalance, we incorporate class weighting and data augmentation techniques. Experimental results demonstrate that the late fusion approach effectively captures patterns of hate speech that may be missed when analyzing a single modality. This highlights the importance of multimodal strategies in enhancing hate speech detection, particularly for low-resource languages.

1 Introduction

The rise of hate speech on social media necessitates automated detection for safer online spaces (Schmidt and Wiegand, 2017). While significant progress has been made in high-resource languages like English, research in Tamil, Malayalam, and Telugu remains limited (Zampieri et al., 2019). The linguistic complexity of Dravidian languages—rich morphology, agglutinative structures, and unique syntax—poses additional NLP challenges (Hegde et al., 2021). Hate speech is prevalent in both text and speech, especially on video-sharing and voice-based platforms (Kumar et al., 2021). Advancements in deep learning and transformer models have enabled more accurate multimodal detection (Kiela et al., 2020).

Dravidian languages suffer from insufficient labeled data, limiting supervised learning (Chakravarthi et al., 2021). Hate speech datasets are highly imbalanced, with fewer hateful instances (Saha et al., 2021), and complex linguistic features like phonetic variations and dialectal differences further challenge text and speech processing (Krishnan et al., 2022). While Muril shows promise for Indian language text processing (Khanuja et al., 2021), speech models like Wav2Vec 2.0 require adaptation for Dravidian languages.

This study introduces a multimodal hate speech detection model integrating Muril for text and Wav2Vec 2.0 for speech, employing a late fusion technique to address these challenges.

2 Related Works

Historically, hate speech detection relied on text-based models like SVMs, Naïve Bayes, and Random Forests (Davidson et al., 2017). Deep learning models, including LSTM, CNNs, and Transformers (BERT, RoBERTa, XLM-R), improved performance, especially in English (Zampieri et al., 2019), but struggle with implicit hate, sarcasm, and multimodal cues.

With the rise of speech-based platforms, self-supervised models like Wav2Vec 2.0, HuBERT, and Whisper have replaced MFCC- and HMM-based methods (Baevski et al., 2020). Multimodal approaches, such as transformers integrating text and vision (Kiela et al., 2020) and late fusion combining text and speech at the logit level (Yin and Zubiaga, 2021), have further enhanced detection.

The HOLD-Telugu shared task (Premjith et al., 2024a) highlighted transformer effectiveness in Telugu code-mixed text. Expanding on this, (Premjith et al., 2024b) demonstrated multimodal advantages in hate speech detection, while (Lal G et al., 2025)

introduced cost-sensitive learning for class imbalance in Dravidian text. Data augmentation techniques, including back-translation, paraphrasing, and synthetic generation, have improved text-based detection (Founta et al., 2018), while speed variation, pitch shifting, and noise injection enhance speech model robustness.

3 Dataset and Preprocessing

We use the DravidianLangTech@NAACL 2025 dataset, a benchmark for multimodal hate speech detection in Tamil, Malayalam, and Telugu. It contains text and speech samples from social media, labeled into five classes—one non-hate and four hate categories. Given the skewed class distribution (Fig. 1), specialized preprocessing and augmentation techniques are applied to improve model robustness.

3.1 Text Preprocessing

The text data is preprocessed using Unicode Normalization for consistency, Unwanted Character Removal to retain only meaningful text, Sentence Splitting and Tokenization for structured segmentation, and Stopword Removal to enhance relevance.

3.2 Speech Preprocessing

Speech preprocessing involves resampling all audio samples to 16 kHz to match Wav2Vec 2.0's default input requirements. Noise reduction is applied using spectral subtraction to remove background interference and enhance speech clarity. Finally, feature extraction is performed directly by Wav2Vec 2.0, which generates raw speech embeddings, eliminating the need for manual feature engineering techniques such as MFCCs or spectrogram analysis.

3.3 Data Augmentation

To improve model generalization, data augmentation techniques were employed separately for speech and text because to the class imbalance in the data set. We used data augmentation approaches to improve the generalization and robustness of the model for both audio and textual input.

3.3.1 Text Data Augmentation

- **Synonym Replacement:** Uses a pre-trained FastText model for contextual synonym substitution.
- **Backtranslation:** Introduces lexical and syntactic diversity via intermediate language translation.

3.3.2 Audio Data Augmentation

- **Gaussian Noise Addition:** Injects noise at varying levels (0.005, 0.01, 0.03) to enhance robustness against distortions.

4 Methodology

4.1 Text-Based Model (Muril)

Muril, a transformer-based model pre-trained on 17 Indian languages, excels in Indian language processing, particularly in Dravidian scripts and low-resource settings, outperforming mBERT and XLM-R. It is optimized for hate speech detection using the DravidianLangTech@NAACL 2025 dataset.

Fine-tuning begins with text preprocessing and tokenization using Muril's subword tokenizer. The tokenized input passes through the Muril encoder to generate contextualized embeddings, which are processed by fully connected layers and a softmax classifier to predict six hate speech classes: Non-Hate, Gender Hate, Political Hate, Religious Hate, Religious Defamation, and Personal Defamation. Training is conducted with a batch size of 32, sequence length of 128, and a 3e-5 learning rate using the AdamW optimizer for 10 epochs, with early stopping based on validation loss. Categorical cross-entropy loss is used to optimize the classification problem; the loss function is provided by:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{1}$$

where y_i represents the ground-truth label, and $\hat{y}_i$ is the predicted probability for class ii.

4.2 Speech-Based Model (Wav2Vec 2.0)

Wav2Vec 2.0 (Baevski et al., 2020) is used for speech-based hate speech detection, learning speech representations directly from raw audio without phonetic transcriptions. Effective in low-resource settings, it handles dialectal variations in Tamil, Malayalam, and Telugu better than traditional MFCC-based classifiers by capturing nuanced phonetic and prosodic features.

The classification pipeline involves preprocessing (Section 3.2), followed by Wav2Vec 2.0 encoding to generate contextualized embeddings, which

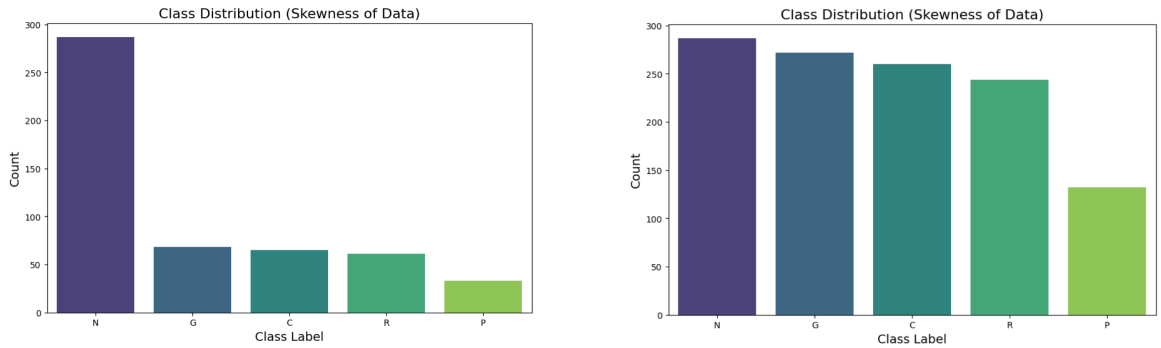


Figure 1: The left side graph depicts the unbalanced data and right side graph is the data distribution after data augmentation

are processed by fully connected layers and classified via softmax. The model is trained independently on the DravidianLangTech@NAACL 2025 dataset using the AdamW optimizer (batch size: 16, learning rate: 2e-5) for 10 epochs, with class weighting to address imbalance. Categorical cross-entropy loss is used for optimization, as in the Muril model.

4.3 Computational Cost

Training was conducted on Google Colab Free with an NVIDIA Tesla T4 GPU (16 GB VRAM), Intel Xeon CPU (2 vCPUs, 2.3 GHz), and 12 GB RAM. Fine-tuning Muril and Wav2Vec 2.0 for Tamil, Malayalam, and Telugu took approximately 1 hour per model over 10 epochs, with GPU utilization reaching 40-60% and peak memory usage of 10 GB.

5 Fusion Techniques

5.1 Early Fusion

Early fusion integrates text and speech features at the representation level by concatenating embeddings from Muril and Wav2Vec 2.0 before classification as shown in Fig. 2. This allows the model to learn cross-modal interactions early in the pipeline. The concatenated feature vector is passed through a shared neural network, which processes both modalities jointly. While early fusion enables deeper multimodal learning, it may introduce modality imbalance, where dominant features, such as text, overshadow weaker ones, such as speech. Additionally, the increased feature dimensionality can lead to overfitting and higher computational costs.

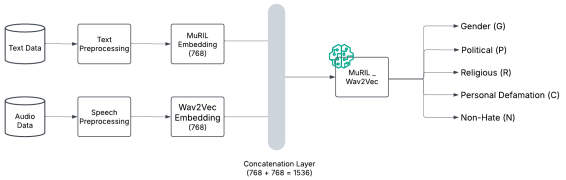


Figure 2: Early Fusion of MuRIL and Wav2Vec for Sentiment Classification

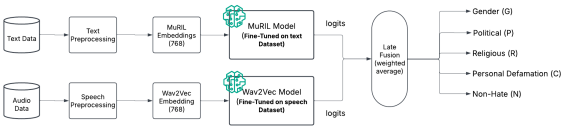


Figure 3: Late Fusion of MuRIL and Wav2Vec for Sentiment Classification

5.2 Late Fusion

Late fusion combines predictions at the decision level rather than merging raw features as shown in Fig. 3. Muril and Wav2Vec 2.0 are trained separately, generating independent class probabilities, P_t for text and P_s for speech. The final classification probability is computed as:

$$P_{\text{final}} = \alpha P_t + (1 - \alpha) P_s \tag{2}$$

where α is a tunable hyperparameter that adjusts the relative contribution of each modality. This approach allows the model to optimize each modality independently before aggregation, reducing the risk of feature redundancy and overfitting.

5.3 Comparison of Fusion Strategies

Early fusion provides stronger cross-modal feature interaction but may suffer from modality dominance and increased computational demands. In contrast, late fusion ensures independent optimization of each modality, offering greater flexibility

in weighting text and speech contributions. By exploring both techniques, we aim to determine the most effective strategy for multimodal hate speech detection.

6 Result

The results of our multimodal hate speech detection model across Tamil, Malayalam, and Telugu demonstrate variations in performance based on fusion strategies and training approaches.

Language	F1 - Train set	F1 - Test set
Tamil	0.79	0.48
Malayalam	0.83	0.51
Telugu	0.73	0.40

Table 1: Train and Test Results using Early fusion

Language	Text		Audio		F1
	Train	Test	Train	Test	Test set
Tamil	0.84	0.74	0.43	0.38	0.71
Malayalam	0.69	0.75	0.65	0.40	0.75
Telugu	0.82	0.35	0.4	0.26	0.17

Table 2: F1 Scores using class weighting (Late Fusion)

Language	Text		Audio		Late
	Train	Test	Train	Test	Fusion
Tamil	0.84	0.69	0.94	0.38	0.70

Table 3: F1 Scores - Augmented data (Late fusion)

6.1 Early Fusion Performance and Overfitting

Early fusion results (Table 1) indicate that while the model achieves relatively high F1-scores on the train set (0.79–0.83), the test set performance drops significantly (0.40–0.51), suggesting overfitting. This is likely due to the absence of explicit regularization techniques such as dropout or weight decay. The model memorizes training patterns but fails to generalize well on unseen data.

6.2 Late Fusion Generalization

Unlike early fusion, late fusion achieves better generalization without explicit regularization. This suggests that independent training of text and audio modalities before aggregation helps mitigate overfitting. Class weighting further balances the contributions of both modalities, leading to improved test performance for Tamil (0.71) and Malayalam (0.75) as shown in (Table 2). However, Telugu's performance remains weak across all modalities. Table 4 highlights that while classes like R and P perform well, G suffers from poor recall (0.30), indicating difficulty in identifying certain instances, which suggests modality-specific challenges.

Class	Precision	Recall	F1-score
C	0.58	0.70	0.64
N	0.56	0.90	0.69
R	0.82	0.90	0.86
P	0.88	0.70	0.78
G	1.00	0.30	0.46
Accuracy		0.70	
Macro Avg	0.77	0.70	0.69
Weighted Avg	0.77	0.70	0.69

Table 4: Classification report of the model trained using the late fusion-class weighting approach.

6.3 Impact of Data Augmentation

Data augmentation (Table 3) improves model robustness, particularly for Tamil, where the test F1-score for text increases to 0.69, and late fusion achieves 0.70. However, augmentation has a minimal effect on Telugu, reinforcing the hypothesis that linguistic characteristics and data sparsity play a larger role in its underperformance.

6.4 Analysis of Telugu's Underperformance

Telugu shows the weakest performance across all models, especially in late fusion (0.17), due to its high phonetic and syntactic diversity, which hinders both text and speech models. Additionally, Wav2Vec 2.0, trained on high-resource languages, struggles with Telugu's unique phonetic structure, leading to lower classification accuracy.

7 Conclusion

This study examines multimodal hate speech detection in Tamil, Malayalam, and Telugu using Muril and Wav2Vec 2.0. Comparing fusion strategies, we find that early fusion enables cross-modal interactions but suffers from overfitting, while late fusion generalizes better by optimizing text and speech models independently.

Class weighting and data augmentation enhance performance, particularly for Tamil and Malayalam, though Telugu remains challenging due to linguistic complexity and data sparsity. Future work will focus on reducing overfitting with regularization techniques, evaluating advanced transformer models, and improving interpretability for better linguistic adaptation.

The implementation of our model, including preprocessing and training scripts, is publicly available at GitHub Repository.

8 Limitations

The class disparity is still a problem in the DravidianLangTech@NAACL 2025 dataset, especially for hate categories related to politics and religion. The lack of pre-trained models for Dravidian languages, along with background noise and accent fluctuation, make speech processing difficult. Because the text-based model performs better than the speech-based model, the modalities' contributions are unbalanced, and late fusion is unable to adequately reflect their complex interconnections. Changes in hate speech patterns and a lack of discourse-level knowledge hinder generalization to real-world contexts.

9 Ethics Statement

This research focuses on improving multimodal hate speech detection while ensuring fairness, transparency, and ethical considerations. We acknowledge the potential for bias in dataset distribution, which may affect classification performance across different hate speech categories. To mitigate this, we incorporate class balancing techniques and assess misclassification trends through error analysis. All data used in this study is publicly available, and no personally identifiable information was processed. While our model aims to enhance online safety, we recognize the risks of false positives and false negatives, which highlight the need for human oversight in real-world applications. We encourage responsible AI deployment and emphasize that this work should not be used to unjustly suppress free speech but rather to foster safer online interactions, particularly in low-resource languages.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Raghavendra Kumar, and E. Sherly. 2021. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages. In *Forum for Information Retrieval Evaluation (FIRE)*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Jayaprakash Hegde, Raghavendra Kumar, Bharathi Raja Chakravarthi, and K. P. Soman. 2021. Classification of offensive language in dravidian languages using deep learning approaches. In *Forum for Information Retrieval Evaluation (FIRE)*.

Simran Khanuja, Sudip Dandapat, Raghavendra Kumar, Sunayana Sitaram, and Monojit Choudhury. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vinayak Goswami, Davide Testuggine, and Peter West. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Prasanna Krishnan, Anand Subramanian, Bharathi Raja Chakravarthi, and K. P. Soman. 2022. Speech recognition in tamil and malayalam using self-supervised learning. In *Proceedings of the ACL Conference on Computational Linguistics*.

Anil Kumar, Ajay Kumar Ojha, Shervin Malmasi, and Marcos Zampieri. 2021. Benchmarking aggression identification in social media for low-resource languages. In *Proceedings of the Workshop on Online Abuse and Harms (ACL)*.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

Prithviraj Saha, Binny Mathew, Narendra Ghanghor, Pawan Goyal, and Animesh Mukherjee. 2021. Hate speech detection in indic languages: A comparative study. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the International Workshop on NLP for Social Media (ACL)*.

Jun Yin and Arkaitz Zubiaga. 2021. Multimodal hate speech detection: A review and open challenges. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL-HLT*.

Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian languages: DravidianLangTech@NAACL 2025

Jyothish Lal G[1], Premjith B[1], Bharathi Raja Chakravarthi[2], Saranya Rajiakodi[3], Bharathi B[4], Rajeswari Natarajan[5], Ratnavel Rajalakshmi[6]
[1]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India,
[2]School of Computer Science, University of Galway, Ireland,
[3]Central University of Tamil Nadu, India, [4]SSN College of Engineering, Tamil Nadu, India,
[5]SASTRA University, India [6]Vellore Institute of Technology, Chennai, Tamil Nadu, India

Abstract

The detection of hate speech in social media platforms is very crucial these days. This is due to its adverse impact on mental health, social harmony, and online safety. This paper presents the overview of the shared task on Multimodal Hate Speech Detection in Dravidian Languages organized as part of DravidianLangTech@NAACL 2025. The task emphasizes detecting hate speech in social media content that combines speech and text. Here, we focus on three low-resource Dravidian languages: Malayalam, Tamil, and Telugu. Participants were required to classify hate speech in three sub-tasks, each corresponding to one of these languages. The dataset was curated by collecting speech and corresponding text from YouTube videos. Various machine learning and deep learning-based models, including transformer-based architectures and multimodal frameworks, were employed by the participants. The submissions were evaluated using the macro F1 score. Experimental results underline the potential of multimodal approaches in advancing hate speech detection for low-resource languages. Team SSNTrio achieved the highest F1 score in Malayalam and Tamil of 0.7511 and 0.7332, respectively. Team lowes scored the best F1 score of 0.3817 in the Telugu sub-task.

1 Introduction

As social networks have become an essential part of modern life, people use it to share their creations, opinions, and daily experiences (Paval et al., 2024). Although it is meant to be a platform for fun and information sharing, some use it to spread hate and profane content (Ben-David and Fernández, 2016). Hate posts usually target specific individuals, groups, or even nations. Moreover, many people use fake profiles to share such content. This harmful behavior has caused mental distress and conflicts among different groups. Addressing this issue requires better detection of hate speech, which is challenging due to the unique language styles used in online content (Schmidt and Wiegand, 2017).

In this context, manually identifying and removing hate speech is the simplest method, but it is a time consuming and tedious process (MacAvaney et al., 2019). Automated methods are becoming more popular because they are faster and more efficient. Most studies focus on detecting hate speech in written text, while progress in videos and images has been made using multimodal datasets. However, detecting hate speech in spoken language and its combination with text has not been explored much due to lack of multimodal datasets, especially for low-resource Dravidian languages (Anilkumar et al., 2024).

A lot of research is being done to recognize hate speech using images, text, and videos (Davidson et al., 2017; Safaya et al., 2020). Most studies currently focus on single modality, especially text. Neverthless, research on speech-based and multimodal approaches are also in growing phase. For example, Abhishek et al. (Anilkumar et al., 2024) used deep learning models to detect hate speech in Hindi and Marathi texts from the HASOC 2021 dataset (Velankar et al., 2021). They showed that transformer models work best, but even simple models with FastText embeddings can give strong results. Similarly, Tashvik Dhamija's (Dhamija et al., 2021) study on English tweets showed that RoBERTa embeddings combined with a decision tree algorithm performed exceptionally well.

A few other studies considered different languages and techniques for detection of hate-speech. For example, Vandan Mujadia et al. (Mujadia et al., 2019) have done hate/offensive content detection in multiple langauges such as Hindi, English, and German. The study showed best results while combining a voting system with ML classifier models. Another work by Joshi et al. (Joshi et al., 2021)

showed that the BERT-based models perform better for Hindi data. Furthermore, Badjatiya et al. showed that a combination of LSTM with gradient-boosted decision trees gave good results. (Badjatiya et al., 2017). In summary, all the aforesaid studies show that advanced algorithms from the ML paradigm can improve the detection of hate speech, irrespective of the language variations.

The Shared Task on Multimodal Hate Speech Detection in Dravidian Languages, held at DravidianLangTech@NAACL 2025, is a step towards advancing the challenges in hate speech detection. This task aims to develop and test methodologies for detecting hate speech in social media using both speech and text.

2 Task Description

This shared task on hate speech detection focuses on Malayalam, Tamil, and Telugu, the three important Dravidian languages. Consequently, the task is divided into three sub-tasks as follows:

- Task 1: Multimodal hate-speech detection in Malayalam
- Task 2: Multimodal hate-speech detection in Tamil
- Task 3: Multimodal hate-speech detection in Telugu

Participants are provided with training data sets that contain multimodal content, including text and speech. The objective is to develop models capable of analyzing these components to predict the appropriate labels for hate speech detection. Model performance will be evaluated using the macro-F1 score, a widely used metric in NLP for classification tasks.

Language	Data	Label	Count	Total
Malayalam	Train	N	406	883
		C	186	
		P	118	
		R	91	
		G	82	
	Test	N	10	50
		C	10	
		P	10	
		R	10	
		G	10	
Tamil	Train	N	287	514
		C	65	
		P	33	
		R	61	
		G	68	
	Test	N	10	50
		C	10	
		P	10	
		R	10	
		G	10	
Telugu	Train	N	198	556
		C	122	
		P	58	
		R	72	
		G	106	
	Test	N	10	50
		C	10	
		P	10	
		R	10	
		G	10	

Table 1: Distribution of the hate speech data in Train and Test sets in Malayalam, Tamil, and Telugu languages. Here, the class labels N, C, P, R and G represent Non hate, Personal defamation, Political hate speech, Religious hate speech and Gender-based hate speech, respectively.

3 Dataset description

We collected our dataset from YouTube videos on channels with more than 50,000 subscribers to ensure wide reach and engagement. Instead of using a predefined list of hate speech terms, we selected videos based on the context of the spoken audio. In particular, we focused on topics that likely spark controversial or polarizing discussions. Examples include debates on the Ram Mandir inauguration, defamation of well-known figures. These topics are chosen for their high engagement and potential to attract hateful comments. By manually reviewing the audio for context and intent, we identified nuanced instances of hate speech without relying on specific keywords. Our study focuses on four types of hate speech, as defined by YouTube's Hate speech policy. They are as follows.

- **Gender-based hate speech (G):** Content targeting individuals based on their gender identity, sexual orientation, or personal relationships.
- **Political hate speech (P):** Negative remarks directed at individuals based on their nationality or political beliefs.
- **Religious hate speech (R):** Hateful content

aimed at specific individuals or communities related to their religion.

- **Personal Defamation (C):** Dehumanizing comments, such as comparisons to animals, diseases, or pests.

For sentences with multiple labels, such as personal defamation and religious content, the context of the video is used for final labeling. Further, we collected Non-hate (N) speech data from motivational videos because they are less likely to include offensive content.

4 Participants Methodology

A total of 134 teams registered for this shared task. However, only 19 teams submitted the results for atleast one sub-tasks. Precisely, 17 teams submitted the results for Malayalam and Tamil sub-task, and 18 teams submitted for Telugu sub-task. Some of the teams submitted multiple runs, of which, best run was taken as their final submission. We evaluated the submissions using the macro F1 score, and then prepared the rank list based on the results. Tables 1, 2 and 3 show the rank lists for the Malayalm, Tamil and Telugu sub-tasks, respectively. The methodologies used by each team are explained in following subsections.

4.1 SSNTrio

The team **"SSNTrio"** extracted speech features from the spectrogram and appended them to the corresponding text transcript and used a language-specific BERT model to complete the classification of hate speech. This approach achieved a macro F1 score of 0.7511, 0.7332, and 0.3758 for Task 1, Task 2, and Task 3, respectively.

4.2 lowes

The team **"lowes"** fine-tuned language open-source BERT models, which were pre-trained on the Dravidian languages by l3cube-pune. This approach achieved a macro F1 score of 0.7367, 0.7225, and 0.3817 for Task 1, Task 2, and Task 3, respectively.

4.3 MNLP

The team **"MNLP"** has used a deep learning-based model. Precisely, the model is fine-tuned for the classification task. This approach reported a macro F1 score of 0.6135, 0.4877, and 0.2184 for Task 1, 2, and 3, respectively.

Team	Macro F1 Score	Rank
SSNTrio (J et al., 2025)	0.7511	1
lowes	0.7367	2
MNLP (Chauhan and Kumar, 2025)	0.6135	3
byteSizedLLM (Manukonda et al., 2025)	0.5831	4
KEC_Tech_Titans	0.5114	5
zerowatts (Shanmugavadivel et al., 2025a)	0.4726	6
gryffindor (Shanmugavadivel et al., 2025d)	0.4725	7
VKG_VELLORE	0.4604	8
NLP_goats	0.4105	9
SSN_IT_SPEECH	0.3726	10
SSN_MMHS (Murali and Sivanaiah, 2025)	0.348	11
The Deathly Hallows (Shanmugavadivel et al., 2025b)	0.3016	12
Bright Red (Shanmugavadivel et al., 2025c)	0.2782	13
cantnlp (Wong and Li, 2025)	0.273	14
Team ML_Forge (Faisal et al., 2025)	0.2005	15
KEC-Elite-Analysts	0.0812	16
deanhthin	0.0758	17

Table 2: Rank list of Malayalam sub-task

Team	Macro F1 Score	Rank
SSNTrio (J et al., 2025)	0.7332	1
lowes	0.7225	2
The Deathly Hallows (Shanmugavadivel et al., 2025b)	0.6438	3
KEC_Tech_Titans	0.5322	4
MNLP (Chauhan and Kumar, 2025)	0.4877	5
KEC-Elite-Analysts	0.4281	6
NLP_goats	0.4049	7
VKG_VELLORE	0.3743	8
cantnlp (Wong and Li, 2025)	0.3186	9
Bright Red (Shanmugavadivel et al., 2025c)	0.3018	10
DLRG (Rajalakshmi et al., 2025)	0.2542	11
zerowatts (Shanmugavadivel et al., 2025a)	0.2432	12
gryffindor (Shanmugavadivel et al., 2025d)	0.2431	13
SSN_IT_SPEECH	0.2099	14
byteSizedLLM (Manukonda et al., 2025)	0.1596	15
Team ML_Forge (Faisal et al., 2025)	0.1346	16
deanhthin	0.0592	17

Table 3: Rank list of Tamil sub-task

Team	Macro F1 Score	Rank
lowes	0.3817	1
SSNTrio (J et al., 2025)	0.3758	2
SemanticCuet Sync_telugu (Hossain et al., 2025)	0.3514	3
VKG_VELLORE	0.3324	4
NLP_goats	0.2991	5
KEC_Tech_Titans	0.2857	6
gryffindor (Shanmugavadivel et al., 2025d)	0.264	7
zerowatts (Shanmugavadivel et al., 2025a)	0.264	8
Bright Red (Shanmugavadivel et al., 2025c)	0.251	9
byteSizedLLM (Manukonda et al., 2025)	0.2271	10
MNLP (Chauhan and Kumar, 2025)	0.2184	11
cantnlp (Wong and Li, 2025)	0.1774	12
SSN_IT_SPEECH	0.1631	13
SSN_MMHS (Murali and Sivanaiah, 2025)	0.1567	14
The Deathly Hallows (Shanmugavadivel et al., 2025b)	0.1559	15
Team ML_Forge (Faisal et al., 2025)	0.1465	16
KEC-Elite-Analysts	0.1326	17
deanhthin	0	18

Table 4: Rank list of Telugu sub-task

4.4 byteSizedLLM

The team **"byteSizedLLM"** has used a customized attention BiLSTM architecture integrated with XLM-RoBERTa base embeddings for textual features. Additionally, The XLM-RoBERTa was fine-tuned to handle the complexities associated with syntax and semantics. For audio features, they used fine-tuned wav2vec2-base multilingual speech embeddings. This approach reported a macro F1 score of 0.5831, 0.1596, and 0.2271 for Task 1, Task 2, and Task 3, respectively.

4.5 KEC_Tech_Titans

The team **"KEC_Tech_Titans"** employed pre-trained language models, including BERT, mBERT, RoBERTa, and XLNet, fine-tuned on task-specific datasets, along with CNN and BiLSTM to capture hierarchical and sequential patterns. They also utilized HAN and HGNN for attention-based feature extraction. For speech, speech-to-text models were integrated with text-based classifiers, and BiLSTM was applied for sequential feature analysis. Here, predictions from speech and text were combined using late fusion. This ensured a balanced classification. This method reported a macro F1 score of 0.5114, 0.5322, and 0.2857 for Task 1, Task 2, and Task 3, respectively.

4.6 zerowatts

The team **"zerowatts"** implemented an audio classification system by extracting acoustic features such as MFCC and spectral contrast, followed by feature normalization and label encoding to prepare data for model input. This method showed a macro F1 score of 0.4726, 0.2432, and 0.264 for Task 1, Task 2, and Task 3, respectively.

4.7 gryffindor

The team **"gryffindor"** developed an audio classification system by extracting acoustic features. The features include normalized MFCC and spectral contrast. This method reported macro F1 scores of 0.4725, 0.2431, and 0.264 for the three tasks in order.

4.8 VKG_VELLORE INSTITUTE OF TECHNOLOGY

The team **"VKG_VELLORE INSTITUTE OF TECHNOLOGY"** adopted a two-stage approach. In the first stage, language-specific models were trained using BERT embeddings. These embeddings are generated from text transcripts with pre-trained models. They have addressed the class imbalance problem through SMOTE and employed a CatBoost classifier for prediction. In the final stage, the whisper model is used for transcribing the audio. Further, the processed text was fed into the corresponding language-specific CatBoost model for classification. This method achieved a macro F1 score of 0.4604, 0.3743, and 0.3324 for the three tasks in order.

4.9 NLP_Goats

The team **"NLP_Goats"** utilized a TF-IDF-based approach for text classification, employing logistic regression to predict class labels. Text preprocessing included tokenization, stopword removal, and bigram generation, followed by TF-IDF vectorization to convert text into numerical features. To address class imbalance, oversampling techniques were applied before training the Logistic Regression model, with performance evaluated using accuracy and other classification metrics. This approach achieved a macro F1 score of 0.4105, 0.4049, and 0.2991 for Task 1, Task 2, and Task 3, respectively.

4.10 SSN_IT_SPEECH

The team **"SSN_IT_SPEECH"** employed a multimodal deep learning approach to detect hate speech by combining features from both audio and text data. Audio features are extracted using MFCC, which captures acoustic characteristics, while text features are derived using TF-IDF to analyze linguistic content. The extracted features are processed by a neural network: audio features pass through dense layers, while text features are handled by LSTM layers, which are well-suited for sequential data. This method leverages both the acoustic and textual properties of speech to achieve robust and nuanced hate speech detection. This approach achieved a macro F1 score of 0.3726, 0.2099, and 0.1631 for Task 1, Task 2, and Task 3, respectively.

4.11 SSN_MMHS

The team **"SSN_MMHS"** employed a multimodal framework for hate speech detection using two encoder-decoder transformer-based pipelines, each incorporating LSTM layers for sequential modeling. The key idea is to enable cross-modality learning by reversing the input modalities across the pipelines. In Pipeline 1, the encoder processes speech features (MFCCs), while the decoder processes text embeddings. Conversely, in Pipeline

2, the encoder processes text embeddings and the decoder processes speech features, fostering better interaction between modalities. The outputs from both pipelines are concatenated to form a unified representation, which is passed through a linear layer followed by softmax for classification. This approach achieved a macro F1 score of 0.348 and 0.1567 for Task 1 and Task 3, respectively.

4.12 The Deathly Hallows

The team **"The Deathly Hallows"** implemented a multimodal approach for classification, combining audio and text features. Audio data was augmented with techniques like noise addition, time-stretching, and pitch-shifting, and MFCCs were extracted for CNN-based classification. Text features were processed using the "*xlm-roberta-large*" model to generate embeddings, followed by an FNN for classification. Both pipelines employed robust architectures with Dropout, BatchNormalization, and Adam optimizer, ensuring accurate and generalized predictions. This approach achieved a macro F1 score of 0.3016, 0.6438, and 0.1559 for Task 1, Task 2, and Task 3, respectively.

4.13 BrightRed

The team **"BrightRed"** preprocessed text and audio data for all three languages and evaluated three models: Random Forest, LSTM, and CNN. Among these, the Random Forest model achieved the highest accuracy. This approach achieved a macro F1 score of 0.2782, 0.3018, and 0.251 for Task 1, Task 2, and Task 3, respectively.

4.14 cantnlp

The team **"cantnlp"** trained the multimodal hate speech classification model using logistic regression by transforming the audio files as melSpectrogram, which implicitly encodes linguistic information as acoustic features. They compared the performance with the text data across multiple statistical language models such as Naive Bayes Classifier for Multinomial Models, Linear Support Vector Machine, Logistic Regression, and Random Forest Classifier. This approach achieved a macro F1 score of 0.273, 0.3186, and 0.1774 for Task 1, Task 2, and Task 3, respectively.

4.15 Team ML_Forge

The team **"Team ML_Forge"** implemented a multimodal training approach, combining text and audio features. Text data was upsampled using back-translation, while audio data was processed at a 16 kHz sampling rate and augmented with variations in sound, pitch, volume, and time-stretching. Missing audio files in the Tamil and Telugu datasets were identified and addressed. Text features were extracted using the mBERT model, and audio features were processed with the wav2vec model, supplemented by MFCC features. The features from both modalities were concatenated and passed through a fully connected layer to generate final predictions. This approach achieved a macro F1 score of 0.2005, 0.1346, and 0.1465 for Task 1, Task 2, and Task 3, respectively.

4.16 KEC-Elite-Analysts

The team **"KEC-Elite-Analysts"** employed a combination of machine learning classifiers, including Random Forest, Support Vector Machine, Naive Bayes, and XGBoost. These models were used to capture diverse patterns from textual and multimodal inputs, leveraging their strengths for effective hate speech classification in underrepresented languages. This approach achieved a macro F1 score of 0.0812, 0.4281, and 0.1326 for Task 1, Task 2, and Task 3, respectively.

4.17 deanhthin

The team **"deanhthin"** utilized an LSTM model to extract text features and a CNN integrated with log-mel spectrograms to extract audio features. These features were then fused using the Tensor Fusion method. This approach achieved a macro F1 score of 0.0758 and 0.0592 for Task 1 and Task 2, respectively.

4.18 DLRG

The team **"DLRG"** used the pre-trained model "*ai4bharat/indic-bert*" for text classification and "*vasista22/whisper-tamil-medium*" for transcription of audio. Precisely, the audio was converted to text using the Whisper model, and the obtained text were classified using the Indic-BERT model. This approach achieved a macro F1 score of 0.2542 for Tamil.

4.19 SemanticCuetSync_Telugu

The team **"SemanticCuetSync_Telugu"** used "*openai/whisper-small*" for audio feature extraction and "*l3cube-pune/telugu-bert-scratch*" for textual feature extraction. The features were combined using a gated fusion approach to perform hate speech

classification in Telugu. The method achieved a macro F1 score of 0.3514.

The majority of the submissions to this shared task centered around transformer-based models and multimodal frameworks. Leading teams such as SSNTrio and lowes leveraged fine-tuned BERT models augmented with speech features such as spectrograms and MFCC. Teams widely used late fusion techniques and attention mechanisms to fuse the features of text and speech data. For instance, byteSizedLLM combined XLM-RoBERTa for text with wav2vec2 for speech, while KEC_Tech_Titans integrated BERT variants with CNNs/BiLSTMs and graph networks. The efficacy of oversampling algorithms such as SMOTE was integrated into the models by some teams to address the class imbalance problem present in the data. Speech-to-text pipelines using Whisper models (DLRG, SemanticCnetSync_Telugu) and acoustic feature extraction (MFCCs, spectral contrast) paired with CNNs/LSTMs (The Deathly Hallows, zerowatts) highlighted the diversity in audio processing. The top-performing approaches showed the efficacy of fine-tuned transformers and multimodal integration, achieving superior macro F1 scores in Malayalam and Tamil, while Telugu posed greater challenges, with lower overall performance. Overall, the submissions reflected a blend of advanced deep learning architectures, traditional NLP techniques, and innovative multimodal strategies tailored to low-resource language contexts.

5 Conclusion

The shared task on Multimodal Hate Speech Detection in Dravidian languages at DravidianLangTech@NAACL 2025 is a platform to address the research in detecting hate speech in low-resource languages such as Malayalam, Tamil, and Telugu. The task highlighted the effectiveness of transformer-based models, particularly fine-tuned language-specific BERT models, and multimodal approaches that integrate both textual and acoustic features. The participation of different teams showcased various methodologies, from advanced deep learning architectures to traditional machine learning techniques, all aimed at addressing the complexities of hate speech detection in Malayalam, Tamil, and Telugu. The creation of a comprehensive, multiclass, multimodal dataset further enriches the resources available for future research. The results underscore the potential of combining textual and vocal features for robust hate speech detection, paving the way for more inclusive and accurate models in the fight against online hate speech.

Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2).

References

Abhishek Anilkumar, Jyothish Lal G, B Premjith, and Bharathi Raja Chakravarthi. 2024. DravLangGuard: A Multimodal Approach for Hate Speech Detection in Dravidian Social Media. In *Speech and Language Technologies for Low-Resource Languages (SPELLL)*, Communications in Computer and Information Science.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Anat Ben-David and Ariadna Matamoros Fernández. 2016. Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain. *International Journal of Communication*, 10:27.

Shraddha Chauhan and Abhinav Kumar. 2025. MNLP@DravidianLangTech 2025: A Deep Multimodal Neural Network for Hate Speech Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Tashvik Dhamija, Anjum, and Rahul Katarya. 2021. Comparative Analysis of Machine Learning and Deep Learning Algorithms for Detection of Online Hate Speech. In *Advances in Mechanical Engineering: Select Proceedings of CAMSE 2020*, pages 509–520. Springer.

Adnan Faisal, Shiti Chowdhury, Sajib Bhattacharjee, Udoy Das, Samia Rahman, Momtazul Arefin Labib, and Hasan Murad. 2025. Team ML_Forge@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Md Sajjad Hossain, Symom Hossain Shohan, Ashraful Islam Paran, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. SemanticCuetSync@DravidianLangTech 2025: Multimodal Fusion for Hate Speech Detection- A Transformer Based Approach with Cross-Modal Attention. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bhuvana J, Mirnalinee T T, Rohan R, Diya Seshan, and Avaneesh Koushik. 2025. SSNTrio @ DravidianLangTech 2025: Hybrid Approach for Hate Speech Detection in Dravidian Languages with Text and Audio Modalities. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ramchandra Joshi, Rushabh Karnavat, Kaustubh Jirapure, and Ravirai Joshi. 2021. Evaluation of Deep Learning Models for Hostility Detection in Hindi Text. In *2021 6th International conference for convergence in technology (I2CT)*, pages 1–5. IEEE.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate Speech Detection: Challenges and Solutions. *PloS one*, 14(8):e0221152.

Durga Prasad Manukonda, Rohith Gowtham Kodali, and Daniel Iglesias. 2025. byteSizedLLM@DravidianLangTech 2025: Multimodal Hate Speech Detection in Malayalam Using Attention-Driven BiLSTM, Malayalam-Topic-BERT, and Fine-Tuned Wav2Vec 2.0. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Vandan Mujadia, Pruthwik Mishra, and Dipti Misra Sharma. 2019. IIIT-Hyderabad at HASOC 2019: Hate Speech Detection. In *FIRE (Working Notes)*, pages 271–278.

Jahnavi Murali and Rajalakshmi Sivanaiah. 2025. SSN_MMHS@DravidianLangTech 2025: A Dual Transformer Approach for Multimodal Hate Speech Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ks Paval, Vishnu Radhakrishnan, Km Krishnan, G Jyothish Lal, and B Premjith. 2024. Multimodal Fusion for Abusive Speech Detection Using Liquid Neural Networks and Convolution Neural Network. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

Ratnavel Rajalakshmi, R Ramesh Kannan, Meetesh Saini, and Bitan Mallik. 2025. DLRG@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Kogilavani Shanmugavadivel, Malliga Subramanian, Naveenram CE, Vishal RS, and Srinesh S. 2025a. KEC_AI_ZEROWATTS@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Vasantharan K, Prethish G A, and Santhosh S. 2025b. The_Deathly_Hallows@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Nishdharani P, Santhiya E, and Yaswanth Raj E. 2025c. KEC_AI_BRIGHTRED@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, ShahidKhan S, Shri Sashmitha S, and Yashica S. 2025d. KEC_AI_GRYFFINDOR@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and Offensive Speech Detection in Hindi and Marathi. *arXiv preprint arXiv:2110.12200*.

Sidney Wong and Andrew Li. 2025. cantnlp@DravidianLangTech2025: A Bag-of-Sounds Approach to Multimodal Hate Speech

Detection. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025

Premjith B[1], K Nandhini[2], Bharathi Raja Chakravarthi[3], Durairaj Thenmozhi[4], Balasubramanian Palani[5], Sajeetha Thavareesan[6], Prasanna Kumar Kumaresan[7],
[1]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India,
[2]School of Mathematics and Computer Sciences, Central University of Tamil Nadu, India,
[3]School of Computer Science, University of Galway, Ireland,
[4]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India,
[5]Indian Institute of Information Technology Kottayam, Kerala, India,
[6]Department of Computing, Eastern University, Sri Lanka,
[7]Data Science Institute, University of Galway, Ireland

Abstract

The detection of AI-generated product reviews is critical due to the increased use of large language models (LLMs) and their capability to generate convincing sentences. The AI-generated reviews can affect the consumers and businesses as they influence the trust and decision-making. This paper presents the overview of the shared task on Detecting AI-generated product reviews in Dravidian Languages" organized as part of DravidianLangTech@NAACL 2025. This task involves two subtasks—one in Malayalam and another in Tamil, both of which are binary classifications where a review is to be classified as human-generated or AI-generated. The dataset was curated by collecting comments from YouTube videos. Various machine learning and deep learning-based models ranging from SVM to transformer-based architectures were employed by the participants.

1 Introduction

Customers consider using the reviews posted on social media and e-commerce platforms to purchase products, read books, and watch entertainment shows such as movies and dramas. The reviews directly influence the economic outcomes of the businesses and create perceptions about them (Luo et al., 2023; Jabeur et al., 2023; Tiwari et al., 2024). The emergence of large language models (LLMs) and their capacity to produce human-like text raises significant concerns about the authenticity of reviews, given their potential to generate both positive and negative feedback. Users can use LLM-based tools to fabricate reviews, mislead customers, and ultimately impact businesses. Therefore, it has become crucial to distinguish between human and AI-generated reviews, a task that is challenging due to the high quality of the text generated by these AI tools. Below are some examples of AI-generated reviews [1]:

- **Example 1**: Love this! Well made, sturdy, and very comfortable. I love it!Very pretty
- **Example 2**: It's very hard to get a pair of these pants in the stores. They are a bit too small and too tight.

It is crucial to devise methodologies to detect AI-generated product reviews to ensure the authenticity of the online customer feedback. Numerous studies have been reported regarding the detection of AI-generated product reviews, ranging from traditional classifiers to advanced neural network architectures. (Wani et al., 2024) achieved an accuracy of 98.46% using a hybrid BiLSTM-Word2Vec model, while (Lee et al., 2022) and (Venugopala et al., 2024) used machine learning classifiers such as random forest and support vector machine (SVM). Though neural models (BiLSTM, GRU) often outperform traditional methods in NLP tasks, the absence of standardized benchmarks complicates direct comparisons. High accuracy claims, such as (Wani et al., 2024)'s 98.46%, risk overfitting concerns unless validated on diverse, real-world datasets. (Fayaz et al., 2020) address robustness via an ensemble model, yet the efficacy of majority voting hinges on the diversity and strength of base classifiers, a nuance not thoroughly explored. A critical issue lies in dataset heterogeneity. Studies use disparate sources (Amazon reviews, restaurant reviews) of varying sizes. Most datasets are English-only, limiting insights into

[1]https://osf.io/tyue9/

multilingual detection. (Gambetti and Han, 2023) introduce a GPT-based model focused on linguistic complexity, a promising feature absent in other works, but their analysis lacks cross-validation against existing methods. The reliance on static embeddings like Word2Vec restricted the performance, as newer embeddings better capture semantic nuances. While these studies demonstrate technical proficiency, real-world applicability remains underexplored. (Gambetti and Han, 2023)'s linguistic complexity analysis offers a novel direction but requires integration with behavioral or metadata features for holistic detection. While current methods show promise, their fragmentation across datasets and techniques underscores the need for cohesive frameworks. In addition, gold-standard corpora and models are unavailable for low-resource languages, such as Dravidian languages. Bridging these gaps will be essential to developing robust, adaptable solutions for AI-generated review detection.

The shared task on "Detecting AI-generated product reviews in Dravidian languages: DravidianLangTech@NAACL 2025" offers an avenue for the researchers to differentiate the human and AI-generated reviews in Malayalam and Tamil languages. This is the first instance of conducting a shared task specifically for these two languages. The shared task has two tasks—one in Malayalam and another in Tamil. This shared task introduces a novel corpus in the Malayalam and Tamil languages, which was curated by collecting comments received for YouTube review videos. We considered reviews written in Malayalam and Tamil scripts while excluding those written in Latin letters to maintain the consistency in scripts.

2 Task Description

There are two subtasks in this shared task:

- Task 1: Detecting AI-generated reviews in Malayalam
- Task 2: Detecting AI-generated reviews in Tamil

In both tasks, the objective is to classify a given review into human and AI categories.

3 Dataset Description

The dataset is prepared in Malayalam and Tamil languages. We created the dataset by maintaining an equal number of data points in the human and AI classes in both languages. This artificially created balance prevents the machine learning models from biasing toward any specific class. However, in real-world scenarios, AI-generated reviews are generally less frequent than human-written ones. Imbalanced data collected from real-world scenarios reflects the practical challenges pertaining to this task.

We prepared the dataset by considering reviews collected from various YouTube channels. Different channels were considered for preparing the training and testing data to ensure separate training and testing data distribution. We didn't consider the reviews from e-commerce platforms, and therefore we didn't conduct a cross-domain generalization test.

AI-generated reviews were found to be linguistically less complex than human-written ones. However, distinguishing between human and AI-generated reviews became difficult when AI tools were fine-tuned to mimic human writing styles effectively. Since AI-generated text can be paraphrased or structured to resemble authentic reviews, models had to rely on subtle textual features, making classification more challenging. A major challenge faced during the data collection phase was to maintain an equal number of human and AI-generated reviews. We iterated the AI review generation process multiple times to achieve class balance. We tried to reduce the bias toward any class or product by including reviews related to different product categories collected from different YouTube channels. However, we haven't addressed the problems pertaining to the gender and racial biases.

As generative AI models continue to evolve, their ability to mimic human writing will improve, which makes the differentiation harder. This dataset is the first in Malayalam and Tamil related to this task and provides a benchmark for current models. By expanding the dataset and collecting reviews from e-commerce platforms, we can enhance its generalizability and reduce its various biases.

3.1 Malayalam data

We gathered reviews from YouTube channels by categorizing various product types. To keep the train and test distribution different, we considered different sets of categories for training and testing data, and these data were collected from different YouTube channels. Table 1 shows the list of cate-

Data	Product Categories
Train	Facewash, Referral products, Dress, Makeup products, Meesho products, Furniture, Mobile phone, Movie, TV, Car, Bike, Apple Vision Pro, Laptop, Electronic gadgets, Airpod, BSNL, Internet connection, Food
Test	Food, Books, Car, Bike, Movie, Credit card, Insurance

Table 1: Categories of products used for creating the corpus

Language	Data	Class	Count	Total
Malayalam	Train	Human	400	800
		AI	400	
	Test	Human	100	200
		AI	100	
Tamil	Train	Human	403	808
		AI	405	
	Test	Human	52	100
		AI	48	

Table 2: Distribution of the data in Train and Test datasets in Malayalam and Tamil languages

gories used in training and testing data.

The flow of the dataset creation process is illustrated in Figure 1. Here, we considered two classes: human and AI. The human class includes reviews written by humans, while the AI class includes reviews generated using AI tools. Initially, the comments were collected from the YouTube videos discussing the products mentioned in Table 1. While collecting the data, we made sure that the selected comments contained only Malayalam characters. We removed all other comments during the preprocessing step. Additionally, we eliminated all emojis from the reviews. We prepared the human class data by using these reviews. We used ChatGPT to generate AI-based reviews using the approach put forward by (Xylogiannopoulos et al., 2024). During this process, we provided ChatGPT with human-written reviews and instructed it to generate 20 similar reviews in Malayalam. From the AI-generated reviews, we removed reviews containing less than 5 words and created the AI-generated review corpus. If the number of AI-generated reviews is less, we repeat the instruction until both human and AI classes have an equal number of samples. We ensured that the number of data samples for both the human and AI classes was equal.

3.2 Tamil data

The product feedback template is created using Google Forms with 20 products such as soap, shampoo, hair oil, footwear, wristwatches, mobiles, cosmetics, clothing, handbags, bikes, laptops, and so on. The individuals are advised to submit their feedback experience, both positive and negative, in Tamil without using any AI tools. We share the same template with another set of individuals and advise them to utilize various AI tools such as ChatGPT, Julius, Gemini, among others. We check user reviews for duplicates and tag it with appropriate labels such as "human written" or "AI generated". Figure 2 shows the process of Tamil dataset creation.

Table 2 explains the distribution of train and test data used in Malayalam and Tamil tasks.

4 Methodologies used in the Submissions

A total of 130 teams registered for this shared task. However, only 33 teams submitted the results in Malayalam, and 37 teams submitted them in Tamil. Some of the teams submitted multiple runs. We evaluated the submissions using the macro F1 score, and then prepared the rank list based on the results. Tables 3 and 4 show the rank lists for the Malayalam and Tamil tasks, respectively.

The descriptions of the systems used by the participating teams are given below.

4.1 Nitiz

The team implemented the multilingual AI-generated text detection model using the IndicBERT transformer, which employs a multimodal approach with cultural, syntactic, and semantic feature projections to capture nuanced linguistic characteristics in Malayalam and Tamil.

4.2 byteSizedLLM

The team used a hybrid methodology, combining a customized BiLSTM network with a fine-tuned XLM-RoBERTa base model. The XLM-RoBERTa model was fine-tuned using Masked Language Modeling on a subset of the AI4Bharath dataset, enhancing its multilingual understanding. The dataset included original, fully transliterated, and partially transliterated data, allowing the model to learn robust cross-lingual representations and adapt to varying transliteration patterns. The BiLSTM layer fur-

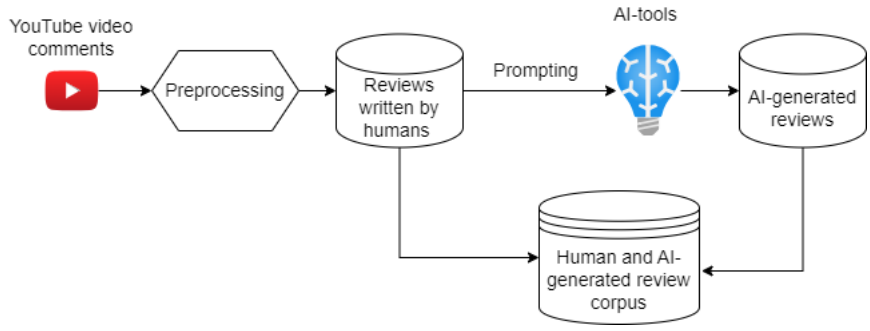


Figure 1: A block diagram explaining the process of creating the Malayalam dataset for the shared task.

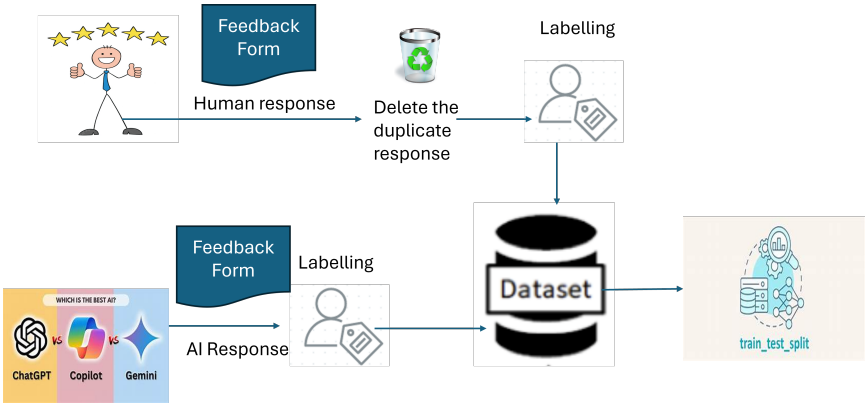


Figure 2: A block diagram explaining the process of creating the Tamil dataset for the shared task.

ther captures sequential dependencies, making it effective for multilingual tasks.

4.3 Girma

The team used Term Frequency-Inverse Document Frequency (TF-IDF) and logistic regression classifier for building the model. The other model proposed by this team had a BERT model trained using Dravidian language data, which outperformed the machine learning model trained using the TF-IDF features.

4.4 InnovateX

The proposed models address distinguishing AI-generated from human-written product reviews in Tamil and Malayalam using SVM, Logistic Regression (LR), and BERT-based transformers. Preprocessing involved text cleaning, tokenization, and label encoding. TF-IDF features (unigrams and bigrams) were used for SVM and LR, while BERT was fine-tuned for contextual understanding.

4.5 Cuet_Absolute_Zero

This team conducted experiments using machine learning models (random forest, support vector machine, decision tree and XGBoost), deep learning models (RNN, GRU, LSTM and BiLSTM) and transformer-based models. The authors increased the number of data points in each class by augmenting new data generated using backtranslation approach.

4.6 CODEGEEK

The model classifies Tamil text as AI-generated or human-generated using a combination of transformer embeddings and a Random Forest classifier. The model uses a pre-trained multilingual BERT model and tokenizer to generate high-dimensional numerical representations of the text's semantic and contextual meaning. These embeddings are used to train a Random Forest Classifier, which learns to classify text based on these embeddings. This combination of deep learning and traditional machine learning ensures effective classification for complex multilingual tasks.

4.7 CIC-NLP

They fine-tuned the XLM-RoBERTa model for text classification tasks in Malayalam and Tamil. Datasets were loaded, preprocessed, and tokenized

Team	F1-score	Rank
KaamKro	0.9199	1
Nitiz	0.9150	2
Three_Musketeers	0.9150	2
SSNTrio (J et al., 2025)	0.9147	3
byteSizedLLM (Kodali et al., 2025)	0.9000	4
Lowes	0.9000	4
CUET_NLP_FiniteInfinity (Hasan et al., 2025)	0.8999	5
TeamVision (S R et al., 2025)	0.8999	5
Necto (Dhasan, 2025)	0.8997	6
Cuet_Absolute_Zero-SIDRATUL	0.8996	7
MUNTAHA (Bijoy et al., 2025)		
Cuet_Absolute_Zero_run-Anindo Barua bijoy	0.8996	7
RATHAN (Thevakumar and Thevakumar, 2025)	0.8994	8
AnalysisArchitects (Jayaraman et al., 2025)	0.8850	9
CIC-NLP (Achamaleh et al., 2025)	0.8849	10
Girma (Bade et al., 2025)	0.8849	10
the_deathly_hallows (Shanmugavadivel et al., 2025)	0.8797	11
CODEGEEK	0.8748	12
MNLP	0.8550	13
AIstudent	0.8350	14
Friends	0.8298	15
Team_Risers (P et al., 2025)	0.8150	16
VKG	0.7834	17
AiMNLP (De and Vats, 2025)	0.7345	18
CUET_NetworkSociety (Aftahee et al., 2025)	0.7287	19
LinguAIsts	0.7100	20
NLP_goats (V K et al., 2025)	0.6849	21
KECLinguAIsts (Subramanian et al., 2025)	0.6697	22
InnovateX (A et al., 2025)	0.6449	23
powerrangers	0.6348	24
VRCLC	0.6310	25
SemanticCuetSync	0.5713	26
Miracle_makers	0.3333	27
HibiscusBots-CIOL	0.1299	28

Table 3: Ranklist of Malayalam sub-task

Team	F1-score	Rank
KEC_AI_NLP	0.9700	1
CUET_NLP_FiniteInfinity (Hasan et al., 2025)	0.9700	1
CIC-NLP (Achamaleh et al., 2025)	0.9600	2
KaamKro	0.9500	3
KEC-Elite-Analysts	0.9499	4
byteSizedLLM (Kodali et al., 2025)	0.9400	5
Nitiz - StarAtNyte	0.9300	6
VKG	0.9299	7
the_deathly_hallows (Shanmugavadivel et al., 2025)	0.9298	8
Team_Risers (P et al., 2025)	0.9197	9
NLP_goats (V K et al., 2025)	0.9099	10
Girma (Bade et al., 2025)	0.8998	11
Three_Musketeers	0.8900	12
AnalysisArchitects (Jayaraman et al., 2025)	0.8800	13
CODEGEEK	0.8678	14
InnovateX (A et al., 2025)	0.8600	15
KECLinguAIsts (Subramanian et al., 2025)	0.8598	16
RATHAN (Thevakumar and Thevakumar, 2025)	0.8368	17
CUET_NetworkSociety (Aftahee et al., 2025)	0.8182	18
AIstudent	0.8140	19
AiMNLP (De and Vats, 2025)	0.7287	20
Lowes	0.7083	21
powerrangers	0.6981	22
Friends	0.6834	23
HibiscusBots-CIOL	0.6745	24
Necto (Dhasan, 2025)	0.6745	24
LinguAIsts	0.6516	25
MNLP	0.6511	26
CUET-NLP_Big_O	0.6419	27
Cuet_Absolute_Zero_run - Anindo Barua bijoy	0.6311	28
Cuet_Absolute_Zero-SIDRATUL	0.6311	28
MUNTAHA (Bijoy et al., 2025)		
SSNTrio (J et al., 2025)	0.5989	29
TeamVision (S R et al., 2025)	0.5586	30
SemanticCuetSync	0.4857	31
Miracle_makers	0.3243	32

Table 4: Ranklist of Tamil sub-task

using the XLM-RoBERTa tokenizer, with labels mapped to numerical codes. The data was split into training and testing sets, and the model was trained using the Hugging Face Trainer API with parameters such as learning rate, batch size, and evaluation strategy. Post-training, we evaluated the model using metrics like accuracy, F1-score, confusion matrices, and ROC curves on the development dataset. Prediction CSV files for the test sets were saved for submission.

4.8 CUET_NetworkSociety

The team used a streamlined machine learning pipeline based on the DistilBERT model, which involved data preprocessing, tokenization, and model training. The preprocessing involved cleaning text to remove HTML tags, punctuation, and numbers, and normalizing whitespace. The tokenized text was then converted into a model-ready format using DistilBERT's tokenizer, ensuring maximum sequence length. The training phase involved splitting data into training and validation sets, with the model trained to maximize the F1 score.

4.9 KECLinguAIsts

The team preprocessed input data by cleaning, removing unwanted characters, and tokenizing reviews. They used TF-IDF vectorization to convert text into numerical features, capturing word importance in each language context. The training dataset was split into 80% for training and 20% for testing. For Tamil, they used Logistic Regression, Random Forest, and XG Boost, while for Malayalam, they used Logistic Regression, MNB, and SVM, each chosen for its ability to handle text data.

4.10 VKG

The proposed system employed the mBERT model (bert-base-multilingual-cased configuration) trained from scratch to classify Tamil and Malayalam product reviews as either human-written or AI-generated. For Tamil reviews, the model achieved a test accuracy of 98.77% and an F1-score of 0.99 for both classes after 5 epochs of training, demonstrating the potential of training multilingual models from scratch for this task, even with lim-

ited data. For Malayalam reviews, the model was trained for 8 epochs.

4.11 LinguAIsts

In this work, initially to preprocess the dataset, labels were encoded into binary values (1 for AI, 0 for humans). Each review was tokenized using BERT, and contextual embeddings were extracted using the [CLS] token, which captures the text's overall semantic meaning. A Support Vector Machine (SVM) classifier with a linear kernel used these embeddings as input features. 80% of the dataset was used to train the model, while the remaining 20% was used for evaluation.

4.12 Team_Risers

This team used a pre-trained language model fine-tuned specifically for Dravidian languages. This method involved preprocessing a custom dataset for compatibility with the model, which included text cleaning, tokenization, and encoding. The model was then trained on the dataset to adapt to the nuances of the Dravidian languages to correctly classify reviews as human-written or AI-generated.

4.13 Three_Musketeers

The team employed a combination of multilingual transformer models to classify AI-generated and human-generated text in both Malayalam and Tamil datasets. Specifically, for the Malayalam dataset, they utilized XLM-RoBERTa-Large, mBERT (Multilingual BERT), and IndicBERT to leverage their multilingual capabilities and contextual understanding of Indian languages. These models were fine-tuned on the dataset to optimize performance metrics such as F1-score and accuracy. For the Tamil dataset, they used mBERT due to its robust multilingual capabilities and proven effectiveness in handling diverse linguistic structures.

4.14 powerrangers

This team used K-Nearest Neighbor classifier for classifying the product reviews into human and ai categories.

4.15 NLP_goats

In this submission, they applied a machine learning approach for text classification by first preprocessing the text data, which involved removing punctuation, numbers, extra spaces, and converting text to lowercase. The processed text was then transformed into numerical features using TF-IDF with character-level n-grams (unigrams and bigrams), which helps capture important features, especially for languages such as Malayalam. A Logistic Regression classifier was trained on the transformed data, and the model's performance was evaluated using metrics such as accuracy, F1 score, precision, and recall, providing a comprehensive assessment of its effectiveness in classifying the text into predefined categories.

4.16 CUET-NLP_Big_O

The team utilized the BiLSTM+CNN model, which integrates convolutional and bidirectional recurrent layers for text classification. The model starts with an embedding layer with a vocabulary size of 10,000 and dimension of 128, then uses a Conv1D layer with 128 filters and a kernel size of 5. It captures contextual dependencies using a Bidirectional LSTM with 64 units per direction. The model refines features and ensures precise classification using a dense layer and softmax layer.

4.17 Rathan

The study used a pretrained multi-model ensemble approach for classification, using mT5-small, XLM-RoBERTa-base, Sentence-Transformers, and IndicBERTv2-MLM-only as feature extraction models. A dense neural network was trained on the extracted features for classification. A weighted averaging ensemble was used to combine the predictions from these models, with softmax probabilities weighted and averaged based on individual performances on the validation set. The final prediction was determined by selecting the class with the highest combined probability.

4.18 CIC-NLP

The team fine-tuned the XLM-RoBERTa model for text classification tasks in Malayalam and Tamil. Datasets were loaded, preprocessed, and tokenized using the XLM-RoBERTa tokenizer, with labels mapped to numerical codes. The data was split into training and testing sets, and the model was trained using the Hugging Face Trainer API with parameters such as learning rate, batch size, and evaluation strategy.

4.19 TeamVision

This team experimented with models like Bert, Naive Bayes, Random Forest, KNN, LSTM and Decision Tree combined with feature extraction

methods such as Bag of Words, TF-IDF, Count Vectorization, and n-grams. They identified BERT as the most accurate model for detecting AI-generated text in Tamil and Malayalam product reviews.

4.20 CUET_NLP_FiniteInfinity

This team employed Sarvam-1 and Gemma-2-2B, two advanced language models with capabilities in Tamil and Malayalam, among other languages.

4.21 HibiscusBots-CIOL

In this work, the team used language-specific models for each language to encode the text and obtain text embeddings. Additionally, they incorporated general Indic language embeddings to handle any cross-lingual nuances. These embeddings were then passed through a Multi-Layer Perceptron (MLP) for training, which facilitated sentiment prediction. They adopted an adaptive modeling approach, actively tracking the best model throughout the training process using the highest F1 score, and ultimately used the best-performing model for making predictions on the test data.

4.22 SSNTrio

In this work, the team upsampled teh data to eliminate class imbalance. BERT model was used for tokenization and used Tamil BERT and Malayalam BERT for classification.

4.23 AnalysisArchitects

They used SVM, IndiaBERT and ALBERT models to classify the task after encoding labels and vectorizing the text.

4.24 Lowes

This team used two BERT-based models such as multilingual BERT and L3Cube's monolingual BERT. In addition, then authors used a GPT-2 model with a causal language modeling (CLM) objective.

4.25 AiMNLP

This team proposed three models for this task: BERT embedding-based models, CNN+BiLSTM hybrid model and machine learning and machine learning ensemble models. BERT embedding-based models acheieved the highest performance score in both tasks.

4.26 KEC-Elite-Analysts

The team utilized a diverse set of models, including both traditional machine learning algorithms (e.g., Logistic Regression, Naive Bayes, SVM, Random Forest, Gradient Boosting) and deep learning approaches (e.g., HAN, DAN, mBERT, RoBERTa, ALBERT). This combination allowed us to compare and integrate the strengths of different techniques for effective classification.

4.27 Miracle_makers

This method leverages advanced NLP techniques, including preprocessing for Tamil text, feature extraction using embeddings (TF-IDF, GloVe, Word2Vec, BERT), and transfer learning with attention-based transformers (mBERT, RoBERTa). A fine-tuned binary classifier distinguishes AI-generated and human-written reviews, evaluated using metrics like accuracy, F1-score, and macro F1-score for robust detection.

4.28 The Deathly Hallows

In this work, the team implemented a deep learning-based approach to determine whether a given text was written by an AI or a human. They preprocessed the Tamil and Malayalam text data by normalizing, tokenizing, and removing stopwords to enhance feature extraction. For Tamil, Advertools was used to extract stopwords, while for Malayalam, they created a custom stopword list. After preprocessing, they used a pre-trained transformer model, such as BERT, to generate embeddings for the input text. These embeddings were then passed through a neural network for classification, where the model was trained to predict if the text was AI-generated or human-written.

4.29 SemanticCuetSync

This team used the Llama 3.2-3B model. At first they used a prompt to let the model know what to do. Then they finetuned the model with the training set. They used 10 epochs for Tamil and 45 epochs for Malayalam. Additionally, we quantized our Llama model to 4 bits. They used the model from Unsloth AI.

4.30 Friends

They used BERT for classification. This model leverages its powerful bidirectional contextual understanding to excel in tasks like natural language understanding and text classification. By incorporating BERT, this system effectively captures

semantic nuances, making it particularly adept at identifying subtle patterns and relationships in language data.

Different teams employed several methodologies to detect AI-generated reviews in Malayalam and Tamil. Transformer-based models, particularly BERT variants, dominated the submissions, leveraging their multilingual capabilities. Moreover, teams used hybrid architectures, such as combining BiLSTM with XLM-RoBERTa or CNN, to capture sequential and local patterns. Submissions based on traditional machine learning classifiers such as SVM, logistic regression, and random forest trained using TF-IDF features provide baselines. Teams used data augmentation techniques, such as back translation, to improve robustness. The models based on multilingual embeddings and ensemble strategies addressed linguistic nuances in the corpus. To summarize, teams built the models using both advanced deep learning and more traditional machine learning methods. However, transformer-based models excelled in identifying reviews generated using AI.

5 Conclusion

The shared task at DravidianLangTech@NAACL 2025 is organized to address the challenges of detecting AI-generated product reviews in Dravidian languages like Malayalam and Tamil. The models submitted by various teams to the shared task demonstrated the efficacy of transformer-based architectures in distinguishing human-written and AI-generated reviews, with top-performing teams achieving macro F1-scores exceeding 0.97 in Tamil and 0.91 in Malayalam. Hybrid models combining BiLSTM, CNN, or ensemble methods were effective in learning sequential and contextual information in the data. The performance of models trained using traditional machine learning classifiers with TF-IDF features lagged behind deep learning approaches, showing the significance of capturing more semantically rich embeddings. The novel dataset curated for this shared task is a significant contribution to the Dravidian language research. However, the artificial class balancing and lack of cross-domain generalization highlight the need for future work to incorporate more real-world characteristics in the data.

Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2), supported in part of Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

Moogambigai A, Pandiarajan D, and Bharathi B. 2025. InnovateX@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Tewodros Achamaleh, Abiola T.O, Lemlem Eyob, Mebiratu¸Mikiyas, and Grigori Sidorov. 2025. CIC-NLP @DravidianLangTech2025: Detecting AI-generated Product Reviews in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sabik Aftahee, Tofayel Ahmmed Babu, MD Musa Kalimullah Ratul, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET_NetworkSociety@DravidianLangTech 2025: A Transformer-based Approach for Detecting AI-Generated Product Reviews in Low-Resource Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Girma Yohannis Bade, Muhammad Tayyab Zamir, Olga Kolesnikova, José Luis Oropeza, Grigori Sidorov, and Alexander Gelbukh. 2025. Girma@DravidianLangTech 2025: Detecting AI Generated Product Reviews. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Anindo Barua Bijoy, Sidratul Muntaha, Momtazul Arefin Labib, Samia Rahman, Udoy Das, and Hasan Murad. 2025. CUET_Absolute_Zero@DravidianLangTech 2025: Detecting Ai-Generated Product Reviews in Malayalam and Tamil Language Using Transformer Models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Somsubhra De and Advait Vats. 2025. AiMNLP@DravidianLangTech2025: Unmask It! AI-Generated Product Review Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language*

Technologies for Dravidian Languages. Association for Computational Linguistics.

Livin Nector Dhasan. 2025. Necto@DravidianLangTech: Fine-tuning Multilingual MiniLM for Text Classification in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Muhammad Fayaz, Atif Khan, Javid Ur Rahman, Abdullah Alharbi, M Irfan Uddin, and Bader Alouffi. 2020. Ensemble Machine Learning Model for Classification of Spam Product Reviews. *Complexity*, 2020(1):8857570.

Alessandro Gambetti and Qiwei Han. 2023. Dissecting AI-Generated Fake Reviews: Detection and Analysis of GPT-Based Restaurant Reviews on Social Media.

Md. Zahid Hasan, Safiul Alam Sarker, MD Musa Kalimullah Ratul, Kawsar Ahmed, and MohammedMoshiul Hoque. 2025. CUET_NLP_FiniteInfinity@DravidianLangTech 2025: Exploring Large Language Models for AI-Generated Product Review Classification in Malayalam. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bhuvana J, Mirnalinee T T, Rohan R, Diya Seshan, and Avaneesh Koushik. 2025. SSNTrio@DravidianLangTech 2025: Identification of AI Generated Content in Dravidian Languages using Transformers. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sami Ben Jabeur, Hossein Ballouk, Wissal Ben Arfi, and Jean-Michel Sahut. 2023. Artificial Intelligence Applications in Fake Review Detection: Bibliometric Analysis and Future Avenues for Research. *Journal of Business Research*, 158:113631.

Abirami Jayaraman, Aruna Devi Shanmugam, Dharunika Sasikumar, and Bharathi B. 2025. AnalysisArchitects@DravidianLangTech 2025: BERT Based Approach For Detecting AI Generated Product Reviews In Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Maharajan Pannakkaran. 2025. AiMNLP@DravidianLangTech2025: Unmask It! AI-Generated Product Review Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Minwoo Lee, Young Ho Song, Lin Li, Kyung Young Lee, and Sung-Byung Yang. 2022. Detecting Fake Reviews with Supervised Machine Learning Algorithms. *The Service Industries Journal*, 42(13-14):1101–1121.

Jiwei Luo, Guofang Nan, Dahui Li, and Yong Tan. 2023. AI-Generated Review Detection. *Available at SSRN 4610727*.

Sai Sathvik P, Muralidhar Palli, Keerthana NNL, Balasubramanian Palani, Jobin Jose, and Siranjeevi Rajamanickam. 2025. TeamRisers@DravidianLangTech 2025: AI-Generated Product Review Detection in Dravidian Languages Using Transformer-Based Embeddings. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shankari S R, Sarumathi P, and Bharathi B. 2025. TeamVision@DravidianLangTech 2025: Detecting AI generated product reviews in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Vasantharan K, Prethish G A, and Vijayakumaran S. 2025. The_Deathly_Hallows@DravidianLangTech 2025: AI Content Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Rojitha R, Mithun Chakravarthy Y, Renusri R V, and Kogilavani Shanmugavadivel. 2025. KECLinguAIsts@DravidianLangTech 2025: Detecting AI-generated Product Reviews in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Jubeerathan Thevakumar and Luheerathan Thevakumar. 2025. RATHAN@DravidianLangTech 2025: Annaparavai- Separate the Authentic Human Reviews from AI-generated one. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shreeji Tiwari, Rohit Sharma, Rishabh Singh Sikarwar, Ghanshyam Prasad Dubey, Nidhi Bajpai, and Smriti Singhatiya. 2024. Detecting AI Generated Content: A Study of Methods and Applications. In *International Conference on Communication and Computational Technologies*, pages 161–176. Springer.

Srihari V K, Vijay Karthick Vaidyanathan, Mugilkrishna D U, and Durairaj Thenmozhi. 2025. NLP_goats@DravidianLangTech 2025: Detecting AI-Written Reviews for Consumer Trust. In *Proceedings of the Fifth Workshop on Speech, Vision, and*

Language Technologies for Dravidian Languages. Association for Computational Linguistics.

PS Venugopala, Amrith R Naik, Nidhish Shettigar, N Vaishnavi, Pranav R Bhat, and Pranesh Kumar Kodi. 2024. Identifying Deceptive AI Reviews: A Machine Learning Approach. In *2024 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, pages 55–59. IEEE.

Mudasir Ahmad Wani, Mohammed ElAffendi, and Kashish Ara Shakil. 2024. AI-Generated Spam Review Detection Framework with Deep Learning Algorithms and Natural Language Processing. *Computers (2073-431X)*, 13(10).

Konstantinos F Xylogiannopoulos, Petros Xanthopoulos, Panagiotis Karampelas, and Georgios A Bakamitsos. 2024. ChatGPT Paraphrased Product Reviews Can Confuse Consumers and Undermine Their Trust in Genuine Reviews. Can You Tell the Difference? *Information Processing & Management*, 61(6):103842.

Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025

Saranya Rajiakodi[1], **Bharathi Raja Chakravarthi**[2], **Shunmuga Priya Muthusamy Chinnan**[2], **Ruba Priyadharshini**[3], **Rajameenakshi J**[4], **Kathiravan P**[1], **Rahul Ponnusamy**[5], **Bhuvaneswari Sivagnanam**[1], **Paul Buitelaar**[5], **Bhavanimeena K**[1], **Jananayagam V**[1], **Kishore Kumar Ponnusamy**[6]

[1]Department of Computer Science, Central University of Tamil Nadu, India.
[2]School of Computer Science, University of Galway, Ireland
[3]Department of Mathematics, Gandhigram Rural Institute -Deemed to be university, India
[4]Department of Social Sciences, Vellore Institute of Technology, Vellore, India
[5]Data Science Institute, University of Galway, Ireland. [6]Digital University of Kerala, Kerala, India.

Abstract

This overview paper presents the findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media, organized as part of DravidianLangTech @NAACL 2025. The task aimed to encourage the development of robust systems to detect abusive content targeting women in Tamil and Malayalam, two low-resource Dravidian languages. Participants were provided with annotated datasets containing abusive and nonabusive text curated from YouTube comments. We present an overview of the approaches and analyse the results of the shared task submissions. We believe the findings presented in this paper will be useful to researchers working in Dravidian language technology.

Disclaimer: This research paper contains offensive/harmful content for research purposes. Viewer discretion is advised.

1 Introduction

As per United Nations, one in three of all women experience Gender Based Violence (GBV) at least once in their lives[1]. Exposure to abusive content on social networks significantly affects people's emotional states (Soral et al., 2023). However, systems for detecting abusive content targeting women have been well developed and widely deployed for English, there is a significant gap in resources and models for Dravidian languages such as Tamil and Malayalam (Caselli et al., 2020; Park and Fung, 2017). The motivation behind conducting this shared task is to address the growing issue of gender-based online harassment, particularly targeting women, in this digital era of social media platforms (Pandey, 2024; Battisti et al., 2024).

This task focuses on identifying abusive text directed at women in YouTube comments. Given a sentence in Tamil or Malayalam, the goal is to contextually analyze the text and determine whether it contains abusive text that specifically targets women.

Tamil and Malayalam, both Dravidian languages rich in literary heritage and agglutinative structures, are widely spoken in Tamil Nadu, Sri Lanka, Kerala, Singapore and Malaysia. This shared task of identifying abusive text in Tamil and Malayalam languages poses challenges such as handling grammatical and spelling errors, managing code-switching between languages, and detecting specific text patterns that target women. This shared task provides an avenue for researchers to classify abusive content in Tamil and Malayalam languages.

Recent growth in multilingual Large Language Models(LLMs), such as LLaMA, mBERT, XLM-R, and IndicBERT, has significantly improved the accuracy of systems for Indian languages (Touvron et al., 2023; Devlin et al., 2019; Kakwani et al., 2020). However, the effectiveness of abusive content detection is highly dependent on the quality of training data. This task utilizes a dataset that is carefully annotated with guidelines in Tamil and Malayalam languages.

2 Related Works

Recent advances in NLP, driven by open source language models, have significantly improved performance on a variety of tasks. However, the detection of abusive content, particularly in Indian languages, is less explored (Mohan et al., 2025; Shunmuga Priya et al., 2022). Languages like Tamil and Malayalam present unique challenges due to their agglutinative nature, frequent code mixing, and the limited availability of large annotated datasets. However, few works have analyzed the ex-

[1]https://www.unwomen.org/en/articles/facts-and-figures/facts-and-figures-ending-violence-against-women

isting challenges in hate speech detection for Indian languages (Mandl et al., 2021). Ponnusamy et al. (2024) has contributed significantly by providing annotated datasets for offensive content detection, laying the foundation for improving NLP models in Dravidian languages.

Subramanian et al. (2022) conducted a comparative study on three variants of transformer models for the detection of hate speech in Tamil. In 2022, Chakravarthi et al. (2022) organized a shared task on the detection of hope speech in Tamil and Malayalam, evaluating system performance using the F1 score. In 2024, another shared task on detecting homophobia and transphobia in social media comments has gained significant attention, addressing the challenges of identifying hateful content in various Indian languages, including Tamil and Malayalam (Chakravarthi et al., 2024). Another critical area of research is code mixing in Dravidian languages, B et al. (2024) work focused on decoding YouTube comments in code mixed Tamil-English and Malayalam-English.

Understanding abusive content involves the analysis of its grammatical structure. Syntactic and semantic ambiguities play an important role in the identification of abusive language (Waseem and Hovy, 2016). This linguistic complexity also complicates the generation of effective word embeddings (Miaschi and Dell'Orletta, 2020), making it challenging to capture the underlying meanings and relationships, especially in abusive language directed at women.

Most of the existing work focuses on general abusive content, but there is a significant gap in detecting gender-specific abuse, especially directed toward women. This shared task aims to fill this gap by encouraging participants to develop models focused on detecting such targeted abuse.

3 Task Description

The shared task challenge was hosted in Codalab [2]. The task's goal is to classify YouTube comments into two categories: Abusive and Non-Abusive in Tamil and Malayalam languages. Participants were provided with:

- Training and Validation Sets: These sets were annotated with labels to allow participants to train and fine-tune their models effectively.
- Testing Set: This set was unlabeled, requiring participants to generate predictions without the aid of ground truth labels, which were reserved for evaluation purposes.

The availability of pre-trained models can help participants address the challenge of linguistic, contextual, and cultural variations by generating meaningful feature representations.

3.1 Dataset Statistics

The dataset was prepared in Tamil and Malayalam, which are low-resource languages. YouTube comments were scraped using targeted queries focused on controversial and sensitive topics where abuse text against women is commonly found. The dataset collection and annotation process has been illustrated in the Figure 1.

The Inter-Annotator Agreement(IAA) was analyzed for both Tamil and Malayalam datasets on Abusive text targeting Women in Social Media. The annotation process involved six annotators, including four Computer Science students and two Social Work students, comprising four females and two males. For Tamil annotation, two Computer Science Students and a Social Work student took part and likewise the same proportion of Malayalam speaking students were involved in the Malayalam dataset annotation process. Majority voting was used to determine the final labels for each instance. The Krippendorff's Alpha value for Tamil annotations was 0.6474, indicating moderate agreement, while Malayalam annotations achieved a Krippendorff's Alpha value of 0.9573, reflecting near-perfect agreement.

Table 1 shows the dataset statistics. In Tamil, the dataset consists of a total of 2,790 samples in training, with 1,424 non abusive samples and 1,366 abusive samples. The average number of words per sample is approximately 14.5 for Tamil.
The Malayalam training corpus consists of a total of 2,933 samples, with 1,531 abusive samples and 1,402 non-abusive samples. The average number of words per sample is approximately 16 for Malayalam. The test set in both languages was used for final evaluation and ranking, providing insights into how well the model generalizes to unseen data and its overall performance. Figures 2 and 3 present sample sentences from the Tamil and Malayalam corpora, respectively.

[2] https://codalab.lisn.upsaclay.fr/competitions/20701

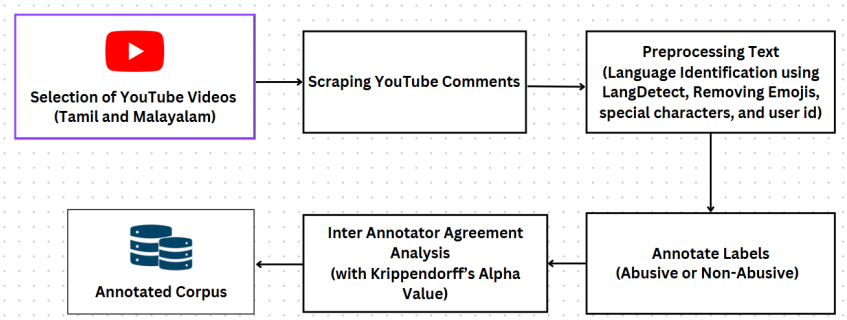


Figure 1: Dataset collection and Annotation Process

Language	Dataset	No of samples	Vocab size
Tamil	Train	2,790	15,863
	Dev/Val	598	4517
	Test	598	4841
Malayalam	Train	2,933	16,344
	Dev/Val	629	4893
	Test	629	4829

Table 1: Dataset Description for Tamil and Malayalam

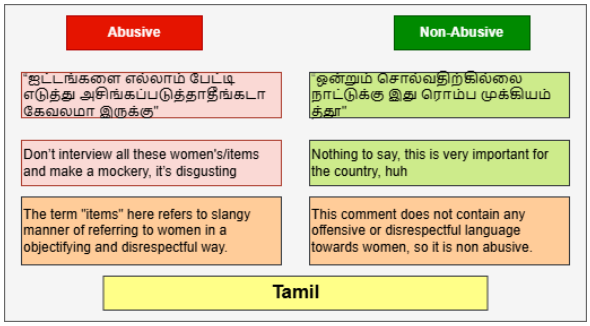


Figure 2: Samples from the corpus for Tamil language

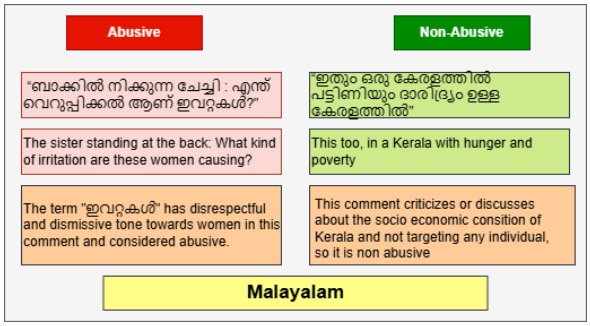


Figure 3: Samples from the corpus for Malayalam language

4 Participant's Methodology

A total of 157 teams registered to participate in this shared task. However, only 37 teams submitted their results for the Tamil, while 35 teams submitted for the Malayalam. This shows a wide variety of methodologies for detecting abusive content targeting women. Transformer-based models such as BERT, mBERT, MuRIL, and XLM-RoBERTa were widely used, with many teams fine-tuning these models to improve their performance on the given dataset. Few teams combined transformer embeddings with traditional machine learning classifiers. For instance, HTMS and KECEmPower used embeddings from models like BERT or TF-IDF and applied Random Forest, SVM, or Logistic Regression to make predictions. Below is a detailed description of each team methodology:

- **ANSR** (Nishanth et al., 2025): Utilized XLM-RoBERTa-XL to extract contextual embeddings from input text. These embeddings were fed into Random Forest (Run 1) and XGBoost (Run 2) classifiers to categorize text as "Abusive" or "Non-Abusive."
- **ARINDASCI**: Employed transformer models, BERT and mBERT, fine-tuning them on Tamil and Malayalam text with tokenization and class label encoding. They applied data augmentation and hyperparameter tuning to address class imbalance problem.
- **Byte-Sized LLM** (Kodali et al., 2025): Hybrid approach was developed, combining attention BiLSTM network with a fine-tuned XLM-RoBERTa base model.
- **Code Crafters**: The team used feature extraction techniques such as Word2Vec, GloVe, and BERT embeddings. Models like Random Forest, LSTM, and pre-trained transformers such as DistilBERT were employed.
- **Courfour IITK** (S et al., 2025): The dataset was normalized, preprocessed to remove incomplete entries, and cleaned to eliminate

punctuation, special characters, and redundant words. The team used NLTK for tokenization and TfidfVectorizer for feature extraction, followed by Random Forest, SVM, and Logistic Regression models.

- **CUET Agile** (Hanif and Rahman, 2025): Tamil BERT and Malayalam BERT were fine-tuned for their respective languages, while IndicBERTv2 was utilized for both. Models were also fine-tuned on unprocessed texts. Each model was trained for 5 epochs using AdamW with a learning rate of 5e-5, with the best validation F1 score determining the optimal epoch.

- **CUET Ignite** (Rahman et al., 2025b): Implemented multilingual BERT with mixed-precision training for faster convergence. The model was fine-tuned using the AdamW optimizer, cross-entropy loss and dynamic learning rate adjustment throughout 15 epochs.

- **CUET NLP FiniteInfinity**: Employed Sarvam-1, Tamil LLaMA 7B Base, and Gemma-2-2B models for fine tuning the model.

- **CUET Novice** (Sayma et al., 2025): Used l3cube-pune/malayalam-bert. To address class imbalance, the team calculated class weights and incorporated them into the loss function to ensure fair training across all classes.

- **CUET Raptors** (Naib et al., 2025): Fine tuning was performed using the PyTorch-based Hugging Face Transformers library, optimizing a single linear classification layer for binary classification. Training was conducted over 5 epochs with a batch size of 16, using the Adam optimizer with weight decay for regularization and binary cross entropy loss.

- **CVF@NITT**: Devleoped IndicBERT and LaBSE embeddings with Bi-GRU(Bidirectional Gated Recurrent Unit), incorporating external knowledge bases.

- **Cyber Protectors** (Rohit et al., 2025): FastText embeddings were generated to create vector representations of the text. Transformed based architecture is utilized for training the model.

- **Falcons**: Fine-tuned MuRIL(Multilingual Representations for Indian Languages) with Adam and binary cross-entropy loss.

- **GS**: Utilized a BERT model for fine tuning.

- **Habiba A, G Aghila** (Habiba A, 2025): Team Habiba A, G Aghila employed a Recurrent Neural Network (RNN) architecture.

- **HTMS** (Harini et al., 2025): The team conducted three runs for the task using machine and deep learning techniques. In Run 1, BERT embeddings were used with a Random Forest classifier, utilizing 5-fold cross-validation. Run 2 combined TF-IDF and BERT embeddings, with dimensionality reduction and fusion, followed by training a Random Forest classifier. In Run 3, TF-IDF embeddings were used with Logistic Regression, employing 5-fold cross-validation for robust evaluation.

- **Hydrangea** (Thirumoorthy et al., 2025): The team used BERT, XLM-RoBERTa, and DistilBERT for three runs on Tamil and Malayalam datasets. Each model was trained for two epochs.

- **Incepto** (Thavarasa et al., 2025): Combined XLM-RoBERTa-base model with a four multi-head attention heads. The extracted features are processed through a deep feed-forward network with layer normalization, and ReLU activation.

- **KEC Tech Titans** (Subramanian et al., 2025a): The team utilized GRU, FastText, and XGBoost models for Tamil, and LSTM, BiLSTM, and XGBoost models for Malayalam in detecting. XGBoost complemented the deep learning models by handling non-linear relationships

- **KECEmPower** (Subramanian et al., 2025b): Applied Logistic Regression, Random Forest, and SVM with TF-IDF embeddings.

- **LinguAIsts** (G et al., 2025): Experimented with SVM, Naïve Bayes, and Logistic Regression using TF-IDF features. Hyperparameter tuning was performed using GridSearchCV.

- **Lexi Logic** (M et al., 2025): The data was cleaned to remove noise, and duplicates. The dataset was balanced using techniques like

oversampling underrepresented classes. The BERT model was then fine tuned on the data.

- **MSM-CUET** (Rahman et al., 2025a): Incorporated MuRIL transformer for malayalam and XLM-RoBERTa for Tamil. Hyperparameter tuning was applied to optimize the training process, and early stopping was introduced to prevent overfitting. Team JAS also fine-tuned the MuRIL model.

- **Necto** (Dhasan, 2025): Fine-tuned a multilingual SBERT model (microsoft/Multilingual-MiniLM-L12-H384) using Tamil and Malayalam data jointly.

- **NLP Goats** (Vaidyanathan et al., 2025): The BERTbase model was fine-tuned with hyperparameters, including a batch size of 16 and a learning rate of 2e-5.

- **NLPopsCIOL** (Nahian et al., 2025): Used custom models specifically trained on Malayalam and Tamil hate speech data to encode the text and extract embeddings. These embeddings were passed through a Multi-Layer Perceptron (MLP) for model training and. A search-based modeling approach was adopted, where the best-performing model was tracked throughout the training process based on the highest F1 score.

- **NomoreHate**: The team employed a fusion model for Malayalam, combining mBERT and Indic-BERT to generate larger and diverse embeddings. For Tamil, the team used a BiLSTM model combined with Indic-BERT, to capture sequential dependencies.

- **ParsePros**: Explored XLM-RoBERTa embeddings integrated with a BiLSTM-based autoencoder.

- **RMKmavericks** (Johnson et al., 2025): Adopted BiLSTM and SVM with TF-IDF for Tamil, and Decision Tree, Random Forest, Multinomial Naïve Bayes for Malayalam.

- **SSN IT NLP** (Maria Nancy et al., 2025): The mBERT model was fine-tuned using cross-entropy loss, with periodic evaluations during training to monitor performance. To address data imbalance, class weights were applied during training to ensure effective learning from both classes.

- **SSN Trio** (T T et al., 2025): Leveraged mBERT and MuRIL for multilingual classification.

- **SSN SQUAD**: This team fine-tuned the mBERT model separately for Malayalam and Tamil using the Hugging Face Trainer. The training process was optimized with hyperparameters, including a learning rate of 2e-5 and a batch size of 16, over five epochs. This team attained a macro F1 score of 0.751 for Tamil and 0.667 for Malayalam.

- **Syndicate IIITK**: Utilized TF-IDF embeddings with an optimized SVM classifier after tokenization and normalization.

- **Techbusters**: Addressed class imbalance with SMOTE(Synthetic Minority Over-sampling Technique). Multiple classifiers, including Random Forest, Naive Bayes, and Decision Tree, were evaluated. Random Forest and Naive Bayes outperformed others.

- **Tewodros**: Implemented Logistic Regression using TF-IDF vectorizer, and hyperparameter tuning using GridSearchCV.

- **Trix**: The preprocessing involved Unicode normalization using the Indic NLP Library, tokenization with Stanza and custom text cleaning to remove non-native characters, stopwords, and standardize spoken-to-written variations. To address class imbalances, upsampling techniques were applied using Scikit-learn's resample function. For feature extraction, CountVectorizer with n-grams was used. Logistic Regression, Multinomial Naive Bayes, and Decision Tree Classifier, were trained and optimized using GridSearchCV.

- **Yadu**: Applied MuRIL with focal loss, label smoothing, to handle class imbalance and a custom multilingual trainer. The model is trained with a batch size of 16, gradient accumulation, a cosine learning rate scheduler and early stopping.

- **YenCS**: ELMo (Embeddings from Language Models) embeddings are then used for feature extraction. The extracted features are fed into a deep learning based classifier.

Table 2: Rank List of Tamil Language

Team Name	F1 Score	RANK
CUET_Agile (Hanif and Rahman, 2025)	0.7883	1
MSM_CUET (Rahman et al., 2025a)	0.7873	2
Incepto (Thavarasa et al., 2025)	0.7864	3
Lexi Logic (M et al., 2025)	0.7824	4
Necto (Dhasan, 2025)	0.7821	5
byteSizedLLM (Kodali et al., 2025)	0.7820	6
CUETNLP FiniteInfinity	0.7767	7
techbusters	0.7721	8
Hydrangea (Thirumoorthy et al., 2025)	0.7708	9
JAS	0.7687	10
SSNTrio (T T et al., 2025)	0.7668	11
Code_Crafters	0.7587	12
NLP_goats (Vaidyanathan et al., 2025)	0.7504	13
KEC TECH TITANS (Subramanian et al., 2025a)	0.7447	14
Cyber_Protectors (Rohit et al., 2025)	0.7356	15
GS	0.7293	16
CUET_Ignite (Rahman et al., 2025b)	0.7224	17
Habiba A ,G Agila (Habiba A, 2025)	0.7207	18
cuetRaptors (Naib et al., 2025)	0.7203	19
ANSR (Nishanth et al., 2025)	0.7201	20
NLPOPSCIOL (Nahian et al., 2025)	0.7039	21
PARSPROSE	0.6998	22
KECEmpower (Subramanian et al., 2025b)	0.6903	23
CoreFour_IIITK (S et al., 2025)	0.6901	24
Syndicate_IIITK	0.6872	25
SSN_IT_NLP (Maria Nancy et al., 2025)	0.6519	26
nomorehate	0.6517	27
YenCS	0.6381	28
Falcons	0.6255	29
LinguAIsts (G et al., 2025)	0.6251	30
RMKMavericks (Johnson et al., 2025)	0.6196	31
CVF@NITT	0.6174	32
TRIX	0.6001	33
VSS	0.5881	34
Yadu	0.5099	35
HTMS (Harini et al., 2025)	0.5007	36
Tewodros	0.3378	37

Table 3: Rank List of Malayalam Language

Team Name	F1 Score	RANK
Habiba A ,G Agila (Habiba A, 2025)	0.7571	1
CUET_Agile (Hanif and Rahman, 2025)	0.7234	2
CUET_Novice (Sayma et al., 2025)	0.7083	3
Incepto (Thavarasa et al., 2025)	0.7058	4
Lexi Logic (M et al., 2025)	0.7001	5
byteSizedLLM (Kodali et al., 2025)	0.6964	6
Necto (Dhasan, 2025)	0.6915	7
ANSR (Nishanth et al., 2025)	0.6901	8
NLP_goats (Vaidyanathan et al., 2025)	0.6843	9
MSM_CUET (Rahman et al., 2025a)	0.6812	10
LinguAIsts (G et al., 2025)	0.6779	11
Hydrangea (Thirumoorthy et al., 2025)	0.6769	12
VSS	0.6757	13
CVF@NITT	0.6701	14
CUETNLP FiniteInfinity	0.6645	15
SSN_IT_NLP (Maria Nancy et al., 2025)	0.6601	16
Cyber_Protectors (Rohit et al., 2025)	0.6518	17
TRIX	0.6501	18
RMKMaveriks (Johnson et al., 2025)	0.6484	19
KECEmpower (Subramanian et al., 2025b)	0.6454	20
techbusters	0.6452	21
NLPopsCIOL (Nahian et al., 2025)	0.6402	22
nomorehate	0.6401	23
Syndicate_IIITK	0.6295	24
ParsePros	0.6201	25
KEC Tech Titans (Subramanian et al., 2025a)	0.6174	26
CoreFour_IIITK (S et al., 2025)	0.6101	27
YenCS	0.5701	28
HTMS (Harini et al., 2025)	0.4947	29
Yadu	0.4801	30
Falcons	0.4772	31
Tewodros	0.3396	32
SSNTrio (T T et al., 2025)	0.3094	33
ARINDASCI	0.2201	34
GS	0.2147	35

5 Results and Discussion

The evaluation metric used was the macro-averaged F1-score, calculated using the scikit-learn library[3].

The Tamil dataset results are shown in table 2. The top performing team, CUET_Agile, achieved a macro F1-score of 0.7883. They have used fine-tuned Tamil BERT model combined with effective optimization strategies. MSM_CUET (0.7873) and Incepto (0.7864) followed closely, utilized multilingual transformer models with fine-tuning techniques. Other notable performers included Lowes, Necto, and ByteSizedLLM, all achieving macro F1-scores above 0.78. These teams demonstrated the effectiveness of pretrained transformer-based models, including MuRIL, XLM-RoBERTa, and BiLSTM.

The Malayalam dataset results are shown in table 3. The top-performing team, Habiba A, G Agila, achieved a macro F1-score of 0.7571, using a Recurrent Neural Network (RNN) approach. Their methodology highlights the potential of traditional deep learning models when combined with effective preprocessing. The second and third positions were claimed by CUET_Agile (0.7234) and CUET_Novice (0.7083), who used fine tuned transformer-based models. Other teams, such as Incepto (0.7058) and Lowes (0.7001), performed well using multilingual pretrained models. The top-performing teams used transformer-based models like BERT, mBERT, MuRIL, and XLM-RoBERTa generally outperformed others. The use of the macro-averaged F1-score and a detailed classification report enabled a fair and comprehensive evaluation of this shared task.

6 Conclusion

This shared task has provided pivotal insights for addressing abuse content targeting women in Dravidian languages. It is evident from the results that the usage of transformer-based models like mBERT and XLM-RoBERTa have out performed the other traditional approaches. In the future, we plan to enhance this task by multiclass problem such as stereotype, bias detection, and gender neutral term analysis. This will enable more contextual analysis for understanding abusive content.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2), supported in part of Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

Premjith B, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024. Findings of the shared task on hate and offensive language detection in Telugu codemixed text (HOLD-Telugu)@DravidianLangTech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55, St. Julian's, Malta. Association for Computational Linguistics.

Michele Battisti, Ilpo Kauppinen, and Britta Rude. 2024. Breaking the silence: The effects of online social movements on gender-based violence. *European Journal of Political Economy*, 85.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of third shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José García-Díaz. 2022. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 378–388, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Livin Nector Dhasan. 2025. Necto@DravidianLangTech: Fine-tuning Multilingual MiniLM for Text Classification in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Dhanyashree G, Kalpana K, Lekhashree A, Arivuchudar K, ARTHI R, Bommineni Sahitya, Pavithra J, and SANDRA JOHNSON. 2025. LinguAIsts@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

DR G AGHILA Habiba A. 2025. DLTC-NITPY@DravidianLangTech 2025 Abusive Code-mixed Text Detection System Targeting Women for Tamil and Malayalam Languages using Deep Learning Technique. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Tareque Md Hanif and Md Rashadur Rahman. 2025. CUET_Agile@DravidianLangTech 2025: Fine-tuning Transformers for Detecting Abusive Text Targeting Women from Tamil and Malayalam Texts. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bachu Naga Sri Harini, Kankipati Venkata Meghana, Kondakindi Supriya, Tara Samiksha, and Premjith B. 2025. HTMS@DravidianLangTech 2025: Fusing TF-IDF and BERT with Dimensionality Reduction for Abusive Language Detection in Tamil and Malayalam. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sandra Johnson, Boomika E, and Lahari P. 2025. RMKMavericks@DravidianLangTech 2025: Tackling Abusive Tamil and Malayalam Text Targeting Women: A Linguistic Approach. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Maharajan Pannakkaran. 2025. byteSizedLLM@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media Using XLM-RoBERTa and Attention-BiLSTM. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Niranjan Kumar M, Pranav Gupta, Billodal Roy, and Souvik Bhattacharyya. 2025. LexiLogic@Dravidian Lang Tech 2025: Detecting Misogynistic Memes and Abusive Tamil and Malayalam Text Targeting Women on Social Media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2021. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '20, page 29–32, New York, NY, USA. Association for Computing Machinery.

C Maria Nancy, Radha N, and Swathika R. 2025. SSN_IT_NLP@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.

Jayanth Mohan, Spandana Reddy Mekapati, Premjith B, Jyothish Lal G, and Bharathi Raja Chakravarthi. 2025. A multimodal approach for hate and offensive content detection in Tamil: From corpus creation to model development. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Abdullah Al Nahian, Mst Rafia Islam, Azmine Toushik Wasi, and Md Manjurul Ahsan. 2025. NLPopsCIOL@DravidianLangTech 2025: Classification of Abusive Tamil and Malayalam Text Targeting Women Using Pre-trained Models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Md. Mubasshir Naib, Md. Saikat Hossain Shohag, Alamgir Hossain, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. cuetRaptors@DravidianLangTech 2025: Transformer-Based Approaches for Detecting Abusive Tamil Text Targeting Women on Social Media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

S. Nishanth, Shruthi Rengarajan, S. Ananthasivan, Burugu Rahul, and S. Sachin Kumar. 2025. ANSR@DravidianLangTech 2025: Detection of abusive tamil and malayalam text targeting women on social media using RoBERTa and XGBoost. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ayush Pandey. 2024. Gender and the legal discourse: Exploring indian laws and cases. *International Journal of Human Rights Law Review*, 3(4):1–41.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada. Association for Computational Linguistics.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Md Mizanur Rahman, Srijita Dhar, Md Mehedi Hasan, and Hasan Murad. 2025a. MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL Based Transformer Models for Detection of Abusive Tamil and Malayalam Text Targeting Women on Social Media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

MD.Mahadi Rahman, Mohammad Minhaj Uddin, and Mohammad Shamsul Arefin. 2025b. CUET_Ignite@DravidianLangTech-NAACL2025: Detection of Abusive Comments in Tamil Text Using Transformer Models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

VP Rohit, Madhav M, Ippatapu Venkata Srichandra, Neethu Mohan, and Sachin Kumar S. 2025. Cyber_Protectors@DravidianLangTech 2025: Abusive Tamil and Malayalam Text Targeting Women on Social Media using FastText. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Varun Balaji S, Bojja Revanth Reddy, Vyshnavi Reddy Battula, Suraj Nagunuri, and Balasubramanian Palani. 2025. CoreFour_IIITK@DravidianLangTech 2025: Abusive Content Detection Against Women Using Machine Learning And Deep Learning Models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Khadiza Sultana Sayma, Farjana Alam Tofa, Md Osama, and Ashim Dey. 2025. CUET_Novice@DravidianLangTech-NAACL2025: Abusive Comment Detection in Malayalam Text Targeting Women on Social Media Using Transformer-Based Models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

MC Shunmuga Priya, D Karthika Renuka, L Ashok Kumar, and Lovelyn Rose S. 2022. Multilingual Low Resource Indian Language Speech Recognition and Spell Correction using Indic BERT. *Sādhanā*, 47.

W. Soral, M. Bilewicz, and M. Winiewski. 2023. Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavious*, 44:136–146.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

Malliga Subramanian, Kogilavani Shanmugavadivel, Deepiga P, Dharshini S, Anandthakumar S, and Praveenkumar C. 2025a. KEC_TECH_TITANS@DravidianLangTech 2025: Abusive Text Detection in Tamil and Malayalam Social Media Comments Using Machine Learning. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Kogilavani Shanmugavadivel, Indhuja V S, Kowshik P, and Jayasurya S. 2025b. KECEmpower@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Mirnalinee T T, J Bhuvana, Avaneesh Koushik, Diya Seshan, and Rohan R. 2025. SSNTrio@DravidianLangTech2025: LLM Based Techniques for Detection of Abusive Text Targeting

Women. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Luxshan Thavarasa, Sivasuthan Sukumar, and Jubeerathan Thevakumar. 2025. Incepto@DravidianLangTech-2025: Detecting Abusive Tamil and Malayalam Text Targeting Women on YouTube. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shanmitha Thirumoorthy, Thenmozhi Durairaj, and Ratnavel Rajalakshmi. 2025. Hydrangea@DravidianLanTech2025: Abusive language Identification from Tamil and Malayalam Text using Transformer Models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Vijay Karthick Vaidyanathan, Srihari V K, and Thenmozhi Durairaj. 2025. NLP_goats@DravidianLangTech 2025: Towards Safer Social Media: Detecting Abusive Language Directed at Women in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025

Bharathi Raja Chakravarthi[1], **Rahul Ponnusamy**[2], **Saranya Rajiakodi**[3], **Shunmuga Priya Muthusamy Chinnan**[1], **Paul Buitelaar**[2], **Bhuvaneswari Sivagnanam**[3], **Anshid Kizhakkeparambil**[4]
[1]School of Computer Science, University of Galway, Ireland
[2]Data Science Institute, University of Galway, Ireland
[3]Department of Computer Science, Central University of Tamil Nadu, India
[4]WMO Imam Gazzali Arts and Science College, Kerala, India

Abstract

The rapid expansion of social media has facilitated communication but also enabled the spread of misogynistic memes, reinforcing gender stereotypes and toxic online environments. Detecting such content is challenging due to the multimodal nature of memes, where meaning emerges from the interplay of text and images. The Misogyny Meme Detection shared task at DravidianLangTech@NAACL 2025 focused on Tamil and Malayalam, encouraging the development of multimodal approaches. With 114 teams registered and 23 submitting predictions, participants leveraged various pretrained language models and vision models through fusion techniques. The best models achieved high macro F1 scores (0.83682 for Tamil, 0.87631 for Malayalam), highlighting the effectiveness of multimodal learning. Despite these advances, challenges such as bias in the data set, class imbalance, and cultural variations persist. Future research should refine multimodal detection methods to improve accuracy and adaptability, fostering safer and more inclusive online spaces.

Disclaimer: This research paper contains offensive/harmful content for research purposes. Viewer discretion is advised.

1 Introduction

The widespread adoption of social media has transformed digital communication, allowing instantaneous sharing of ideas and fostering global connectivity (Singh et al., 2023). However, this evolution has also led to challenges, particularly the proliferation of harmful content such as misogyny memes (Gasparini et al., 2022). These memes, often combining visual and textual elements, propagate gender-based discrimination, perpetuate stereotypes, and contribute to toxic online environments. Their multimodal nature poses significant challenges for automated detection, as the nuanced interplay between images and text frequently conveys implicit and context-dependent meanings (Kumari et al., 2024). Traditional unimodal detection methods often fail to address this complexity, underscoring the need for advanced multimodal analysis techniques to effectively identify and mitigate such content.

Addressing misogyny in online spaces is a critical step toward fostering inclusive and respectful digital environments. Misogyny memes, by embedding discriminatory messages in humor or satire, not only normalize toxic behaviors but also marginalize women and reinforce societal inequalities. Detecting and moderating these memes is particularly challenging in low-resource contexts where annotated datasets and linguistic resources are scarce (Huang et al., 2024). The task requires innovative approaches that integrate textual and visual modalities to capture the implicit biases and indirect messaging characteristic of these memes. By advancing research in this area, it is possible to combat the spread of harmful content and contribute to safer and more equitable online spaces (Rizzi et al., 2024).

To address this issue, we conducted the second shared task on "Misogyny Meme Detection"[1] under the DravidianLangTech@NAACL 2025[2][3] initiative. This shared task focuses on the automatic detection of misogyny in memes across two languages, including Tamil and Malayalam which are low-resourced. The aim is to foster the development of computational models capable of identifying misogynistic content while accounting for linguistic and cultural variations in online communication.

The task is grounded in several objectives:

1. To encourage the creation of state-of-the-art

[1]https://codalab.lisn.upsaclay.fr/competitions/20856
[2]https://sites.google.com/view/dravidianlangtech-2025/
[3]https://2025.naacl.org/

systems for misogyny detection in memes using multimodal approaches.

2. To promote research in low-resource languages, extending the applicability of NLP technologies beyond high-resource settings.

The shared task attracted significant participation from researchers around the world, demonstrating the growing recognition of the need to address misogyny in online spaces. A total of 114 teams registered for the shared task, with 23 teams successfully submitting their predictions, showcasing a diverse range of methodologies that used both textual and visual modalities to enhance multimodal classification performance. The results indicated that the team DLRG_RR achieved the highest macro F1-score (0.83682) for Tamil, while CUET_Novice team obtained the top macro F1-score (0.87631) for Malayalam, emphasizing the effectiveness of multimodal learning. Despite the success of fusion-based models, challenges such as class imbalance, dataset biases, and subtle variations in misogynistic language remain areas for further exploration. Future research should focus on mitigating these biases, addressing cultural nuances, and improving context-dependent understanding to improve the robustness of misogyny meme detection models, contributing to a safer and more inclusive digital space.

2 Related Work

2.1 Misogyny detection

Misogyny detection in online platforms has been a focal point of research, particularly as the internet continues to foster gender-based hate speech. Early efforts include the Evalita 2018 and IberEval 2018 shared tasks, which introduced the Automatic Misogyny Identification (AMI) challenge to detect misogynistic content in English and Italian texts (Fersini et al., 2018). SemEval 2019 extended this focus to multilingual hate speech, addressing misogyny alongside other forms of hate targeting immigrants and emphasizing the detection of aggressive and non-aggressive speech (Basile et al., 2019).

Recent advances have embraced transformer-based models such as BERT and RoBERTa for misogyny detection, leveraging their capability to understand nuanced language. Multilingual models have been particularly effective for tasks involving diverse linguistic contexts (Devlin et al., 2019; Liu et al., 2019). In the multimodal space, datasets such as Facebook Hateful Memes and new multimodal misogyny-specific datasets have encouraged combining textual and visual cues, as seen in approaches utilizing VisualBERT, UNITER, and CLIP for better classification accuracy (Chen et al., 2020; Radford et al., 2021).

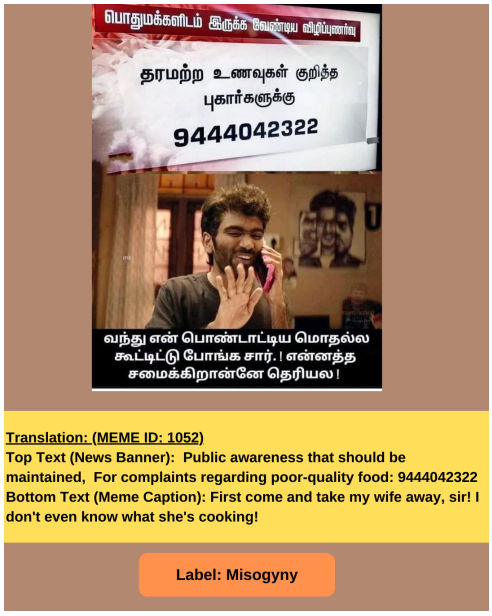
2.2 Multimodal Classification

Multimodal classification methods have become pivotal in tackling tasks involving combined visual and textual data. Traditional approaches used separate pipelines for image and text processing, combining the outputs through simple fusion techniques. Suryawanshi et al. (2020) explored an early fusion technique that combines textual and visual modalities at the embedding level, demonstrating its effectiveness compared to unimodal baselines focused solely on text or images. Koutlis et al. (2023) introduced MemeFier, a deep learning-based framework featuring a dual-stage modality fusion module. This system captured intricate inter-modal connections by integrating feature-level alignment with token-level modality interactions, achieving state-of-the-art results in fine-grained meme classification.

2.3 Related shared tasks

Numerous shared tasks have significantly advanced research on misogyny detection and multimodal content analysis. The Evalita 2018 and IberEval 2018 shared tasks introduced the Automatic Misogyny Identification (AMI) challenge, focusing on detecting misogynistic content in English and Italian (Fersini et al., 2018). These tasks were among the first to address gender-based hate speech systematically. SemEval 2019 expanded the focus to multilingual hate speech detection, including misogyny and hate against immigrants, with distinctions between aggressive and non-aggressive content in English and Spanish (Basile et al., 2019).

Multimodal shared tasks have further enriched the research landscape. The Memotion shared tasks (2020, 2022) targeted the sentiment and emotion analysis of memes, providing valuable benchmarks for multimodal sentiment classification (Sharma et al., 2020; Patwa et al., 2022). The MultiOFF dataset for offensive meme detection highlighted challenges in integrating text and visual modalities for hate speech classification (Suryawanshi et al., 2020). TrollsWithOpinion (2023) introduced a three-level taxonomy for trolling and opinion ma-



(a) Misogyny example from Tamil set

(b) Misogyny example from Malayalam set

Figure 1: Examples from the Tamil and Malayalam sets

nipulation, emphasizing domain-specific opinion manipulation in memes (Suryawanshi et al., 2023).

Our previous shared task, organized as part of LT-EDI@EACL 2024, focused on multitask meme classification with an emphasis on identifying misogynistic and troll content in memes, specifically in Tamil and Malayalam. This task received significant participation, with 52 teams registering and notable submissions. The task A is focused on misogyny meme detection in Tamil and Malayalam language where the top method submitted by MUCS team (Mahesh et al., 2024) achieving macro F1 scores of 0.73 (Tamil) and 0.87 (Malayalam). This effort demonstrated the importance of regional language datasets and the effectiveness of multilingual computational approaches in tackling misogynistic meme detection (Chakravarthi et al., 2024).

3 Task Description

The Shared Task on Misogyny Meme Detection, organized as part of DravidianLangTech@NAACL 2025, aimed to challenge participants to develop advanced multimodal machine learning systems capable of analyzing both textual and visual components of memes. The primary objective of the task was to classify memes as either Misogyny or Non-Misogyny in Tamil and Malayalam languages (Ponnusamy et al., 2024), emphasizing the nuanced intersection of multilingualism and multimodality in the analysis of social media content. The examples from the dataset is show in the Figure 1.

Participants were initially provided with training and development datasets to build and validate their models. Subsequently, a test dataset without labels was released for the final evaluation of the models, which were trained on the previously provided datasets. Submissions was required in a predefined CSV format with their model's prediction using the test set provided for the evaluation. After the release of the results, the labeled test set was shared with participants for personal verification and further analysis.

The datasets included an image folder containing memes in JPG format, accompanied by a CSV file comprising image_id, labels (indicating misogyny or non-misogyny), and transcriptions (text extracted from memes). Participants were expected to adhere to strict submission guidelines and provide their predictions in a predefined format. This shared task served as a platform for exploring cutting-edge methodologies in Natural Language Processing (NLP) and multimodal learning, driving innovation in the analysis of multilingual and multimodal social media content.

4 Participant's Methods

For this shared task, there are total of 114 teams registered, among them we have received a total of 23 submissions where all the participants have

used various types of methodologies to address the task of detecting misogyny and non-misogyny memes in Tamil and Malayalam languages in a multimodal settings from the dataset provided. The participants rank list has been mentioned in the Table 1 for Tamil language and Table 2 for Malayalam language.

- **byteSizedLLM** (Manukonda and Kodali, 2025): This team proposed a multimodal approach for misogynistic meme detection by combining textual and visual features. They fine-tuned the XLM-RoBERTa Base model on Tamil and Malayalam text from the AI4Bharath dataset, using IndicTrans to generate native script, Romanized, and partially transliterated text variations. Text embeddings were mapped to a 768-dimensional space. To align with this, they modified ResNet-50's fully connected layer to extract image features in the same dimension. An Attention-Driven BiLSTM was used to fuse the modalities, capturing sequential dependencies. An attention mechanism followed by a fully connected layer optimized with cross-entropy loss enabled accurate classification.

- **CUET_NetworkSociety**: This team extracted visual features using data augmentation techniques to enhance generalization, employing ResNet18 and EfficientNet-B4. For text processing, they used transformer-based models including IndicBERT for Indian languages, LaBSE for multilingual embeddings, and XLM-RoBERTa for contextual understanding. They applied both classical classifiers—Logistic Regression, SVM, and Random Forest—and deep learning models such as CNN, BiLSTM, and BiLSTM+CNN for textual classification. To integrate text and image modalities, they experimented with concatenation-based fusion (feature-level) and late fusion (prediction-level).

- **Dll5143** (Pattanaik et al., 2025): This team utilized three models—M-CLIP, IndicBERT, and Google's MuRIL—each fine-tuned separately on Tamil and Malayalam meme data to detect misogyny from both images and text. The individual model predictions were combined using majority voting, which enhanced robustness and accuracy by capturing diverse features across modalities and languages. This ensemble approach effectively fused insights from distinct multilingual and multimodal models, improving detection accuracy for memes in Tamil and Malayalam.

- **CUET-NLP_Big_O** (Hossan et al., 2025): This team resized images to 224×224 pixels and normalized them using ImageNet statistics. Tamil and Malayalam texts were tokenized with MuRIL (128 tokens). Visual features were extracted using DenseNet121, EfficientNetB0, ResNet50, and VGG19; textual features used TF-IDF, BoW, and 100-dimensional GloVe embeddings. Features were fused via a fully connected classifier. Training ran for 45 epochs with a 2e-5 learning rate, batch size of 16, and ReduceLROnPlateau scheduler on dual NVIDIA Tesla T4 GPUs, taking 120–150 minutes.

- **Code_Conquerors** (Rao et al., 2025): This team used CLIP model embeddings for image features and BERT-base uncased embeddings for text. For Tamil data, they trained a hybrid model combining ResNet for images and BERT-base uncased for text. For Malayalam data, Vision Transformer replaced ResNet, paired again with BERT-base uncased. In both cases, image and text embeddings were concatenated before training, allowing the model to effectively learn and integrate visual and textual context for improved misogyny detection.

- **Shraddha**: Here, text features were extracted using BiLSTM, and attention mechanisms, while image features were obtained using the ImageNet pre-trained MobileNetV2 model with attention layers. These features were fused for classification, optimized with focal loss and class weights to handle class imbalance.

- **LexiLogic** (M et al., 2025): This study uses L3Cube-Malayalam-BERT and L3Cube-Tamil-BERT for meme categorization and abusive language detection. Data is preprocessed through tokenization and normalization. Fine-tuning uses cross-entropy loss over five epochs at a 2e-5 learning rate. Language-specific embeddings and data augmentation improve performance on low-resource Indian

languages, effectively handling hostile language.

- **One_by_zero** (Chakraborty et al., 2025): They employed CNN, VGG16, and Vision Transformer (ViT) models for visual features extraction, optimizing ViT using the AdamW optimizer and binary cross-entropy loss. BiLSTM, TextCNN, LSTM+CNN, Malayalam BERT, and IndicBERT were used to extract textual features; they were all trained using the Adam optimizer and binary cross-entropy loss. These features were concatenated at a fusion layer and then run through a fully connected classifier that was tuned with Adam and binary cross-entropy loss.

- **teamiic** (Sharma et al., 2025): The XLM-R model was used to handle text data, and the Vision Transformer (ViT) was used to extract features from images. To create a single representation, the embeddings from the two modalities were concatenated. For classification, a proprietary neural network classifier with ReLU activation and a fully connected hidden layer was employed along with cross Entropy Loss and the Adam optimizer.

- **Team_Strikers** (Shanmugavadivel et al., 2025a): This team used LSTM and GRU models to process Tamil-English code-mixed text with TF-IDF, GloVe, and Word2Vec embeddings, while ResNet and EfficientNet CNNs extracted visual features. The CNN-LSTM model combined spatial and sequential learning. Despite challenges with code-mixed input in ResNet-BERT, the GRU-EfficientNet model effectively merged text and visuals.

- **CUET-823** (Mallik et al., 2025): For both text and image inputs, they used text-based augmentation (back-translation via Tamil-English-Tamil and Tamil-Malayalam-Tamil) and image alterations (brightness adjustment, grayscale, posterization) to address class imbalance. They experimented with ViT, ResNet, and EfficientNet for pictures, training for 20 epochs (batch size: 16, learning rate: 1e-4), and optimized mBERT and IndicBERT for text with a 512-token length using the AdamW optimizer (learning rate: 2e-5).

- **CUET_Novice** (Sayma et al., 2025): This team developed a multimodal approach to detect misogyny in Malayalam memes by combining visual and textual features. They used an 8-layer CNN, ResNet-50, Vision Transformer (ViT), and Swin Transformer for visual feature extraction. Text was processed using Malayalam-BERT, generating 768-dimensional embeddings. These were fused with 1024-dimensional Swin Transformer features to create a 1792-dimensional vector. A two-layer neural network with ReLU activation and 0.1 dropout was used for classification. The team trained their model with the AdamW optimizer, binary cross-entropy loss, gradient clipping, a batch size of 16, and a 5e-5 learning rate over five epochs.

- **InnovationEngineers** (Shanmugavadivel et al., 2025b): This team applied padding to text sequences up to a length of 100 before using BERT to extract semantic features. For visual processing, images were resized to 224×224 pixels, normalized to [0, 1], and processed in batches using EfficientNetB0 for feature extraction. They combined both textual and visual features using a Vision-Language Model (VLM) for classification, effectively integrating multimodal data to enhance performance in misogynistic meme detection.

- **Zero_knowledge**: This team employed a multimodal approach, extracting image features using a Convolutional Neural Network (CNN) and processing text through an embedding layer followed by an Long Short-Term Memory (LSTM) to capture sequential and contextual information. Outputs from both branches were concatenated in a fusion layer to integrate visual and textual data. A fully connected dense layer refined the fused features, followed by a sigmoid activation for binary classification. The model was trained using the Adam optimizer with binary cross-entropy loss, ensuring stable and effective convergence for misogynistic meme detection.

- **LinguAIsts** (Arthir et al., 2025): This team used a Support Vector Machine (SVM) with a linear kernel for binary classification, leveraging its effectiveness with textual input vectorized using TF-IDF. To optimize performance,

they applied GridSearchCV to tune hyperparameters such as C, kernel, and gamma, using five-fold cross-validation for robust model selection.

- **Fired_from_NLP** (Chowdhury et al., 2025): This team implemented a multimodal approach, extracting visual features using CNN models like EfficientNetB7, ResNet50, and MobileNetV2, and processing text with Tamil-BERT and Malayalam-BERT. Text preprocessing included padding, truncation, and attention masking via the BERT tokenizer. Images were resized to 224×224 pixels and standardized. Transformer models (mBERT, Indic-BERT, Tamil-BERT, Malayalam-BERT) handled textual feature extraction, while cross-modal attention was used to compute attention scores between text and image features. Outputs were fused using concatenation and the Hadamard product, then passed through dense layers for binary classification. The model was trained with binary cross-entropy loss, early stopping, and a learning rate scheduler.

- **Magma**: For this shared task, They utilized Google's Gemini model to generate dense vector embeddings ('models/embedding-001') which capture the semantic features of Malayalam text. These embeddings were then processed through a Random Forest classifier with 100 estimators, trained on an 80-20 train-test split. For Tamil text classification, They employed a BERT (Bidirectional Encoder Representations from Transformers) model, fine-tuning it specifically for Tamil language processing. The BERT model's bidirectional self-attention mechanisms were adapted to understand Tamil linguistic patterns through the fine-tuning process.

- **CUET-NLP_MP** (Mohiuddin et al., 2025): This team classified Tamil and Malayalam memes using both unimodal and multimodal models. For text, they tested CNN, SVM, Bi-LSTM, mBERT, and XLM-R; for images, they used VGG16, VGG19, ResNet50, Vision Transformer (ViT), and Swin Transformer. The best models for each language and modality were combined for multimodal analysis. Their final model integrated IndicBERT for text and ViT-Base-Patch16-224 for images, with fused embeddings passed through a dense classification layer. Trained over five epochs with a batch size of 16 and a learning rate of 2e-5, this multimodal setup delivered the team's best overall performance.

- **HerWILL** (Preeti et al., 2025): This team adopted a multimodal approach, using language-specific pre-trained models for text encoding: hate-speech-CNERG/malayalam-codemixed-abusive-MuRIL for Malayalam and Tamil-codemixed-abusive-MuRIL for Tamil. For visual features, they used OpenAI's CLIP model (openai/clip-vit-base-patch32) alongside an MLP classifier. They also experimented with a larger vision model (zer0int/CLIP-GmP-ViT-L-14) to explore performance gains. Both early and late fusion strategies were evaluated, along with language models like ai4bharat/IndicBERTv2MLM-only and PosteriorAI/dravida_llama2_7b.

- **vemuri_monisha**: This team combined both the image and text features for this classification task. The image features were extracted using Vision Transformer (ViT), while text features were derived from BERT. These features are then fused and passed through a Random Forest Classifier.

- **SemanticCuetSync**: They fine-tuned the large vision models such as LLaMa 3.2 vision 11b to detect the misogyny content in the dataset provided.

- **DLRG_RR**: This team have utilized the mBERT model to improve the contextual understand in both the Tamil and Malayalam languages.

- **MNLP** (Chauhan and Kumar, 2025): This team used XML-RoBERTa and Byte-Pair Encoding for the extraction of textual features and ViT for the visual features extraction. Then the concatenation based fusion mechanism has been applied and ML models like KNN, SVM, RF, NB and DL models such as LSTM, GRU and Multimodal classifier were used for classification task along with ReLU activation.

5 Results and Discussions

Participants predictions were collected in csv format and evaluated using the macro F1-score, a ro-

Table 1: Rank List of Tamil Language

Team Name	Run	F1 Score	RANK
DLRG_RR	1	0.83682	1
CUET-NLP_Big_O (Hossan et al., 2025)	3	0.81716	2
byteSizedLLM (Manukonda and Kodali, 2025)	3	0.80809	3
CUET-823 (Mallik et al., 2025)	3	0.78120	4
Dll5143 (Pattanaik et al., 2025)	2	0.77591	5
CUET-NLP_MP (Mohiuddin et al., 2025)	1	0.77180	6
CUET_NetworkSociety	1	0.76323	7
MNLP (Chauhan and Kumar, 2025)	2	0.73516	8
LinguAIsts (Arthir et al., 2025)	-	0.71259	9
Shraddha	1	0.70501	10
teamiic (Sharma et al., 2025)	-	0.68830	11
InnovationEngineers (Shanmugavadivel et al., 2025b)	2	0.68782	12
LexiLogic (M et al., 2025)	1	0.68707	13
Fired_from_NLP (Chowdhury et al., 2025)	1	0.67754	14
Code_Conquerors (Rao et al., 2025)	1	0.66142	15
Magma	1	0.65068	16
Team_Strikers (Shanmugavadivel et al., 2025a)	1	0.64776	17
Zero_knowledge	-	0.47801	18
SemanticCuetSync	1	0.40692	19

Table 2: Rank List of Malayalam Language

Team Name	Run	F1 Score	RANK
CUET_Novice (Sayma et al., 2025)	3	0.87631	1
HerWILL (Preeti et al., 2025)	1	0.87483	2
One_by_zero (Chakraborty et al., 2025)	3	0.86658	3
Dll5143 (Pattanaik et al., 2025)	2	0.84927	4
MNLP (Chauhan and Kumar, 2025)	1	0.84237	5
CUET-NLP_MP (Mohiuddin et al., 2025)	1	0.84118	6
teamiic (Sharma et al., 2025)	1	0.84066	7
byteSizedLLM (Manukonda and Kodali, 2025)	1	0.83912	8
CUET-NLP_Big_O (Hossan et al., 2025)	1	0.82531	9
LexiLogic (M et al., 2025)	1	0.80364	10
Fired_from_NLP (Chowdhury et al., 2025)	1	0.80347	11
CUET_NetworkSociety	1	0.80347	12
Code_Conquerors (Rao et al., 2025)	1	0.75649	13
Shraddha	1	0.75467	14
LinguAIsts (Arthir et al., 2025)	-	0.68186	15
Magma	1	0.67552	16
DLRG_RR	1	0.54180	17

bust metric particularly suited for imbalanced classification tasks, ensuring a fair assessment of performance across all classes. Thirty submissions in all were received, and each participant used a different method to identify misogyny and non-misogyny memes in multimodal contexts in Tamil and Malayalam. As per the Tamil findings displayed in Table 1, DLRG_RR obtained the highest rank with a macro F1 score of 0.83682. In order of precedence, CUET-NLP_Big_O came in second with 0.81716, byteSizedLLM with 0.80809, CUET-823 with 0.7812, and Dll5143 with 0.77591. CUET_Novice topped the Malayalam findings in Table 2 with an exceptional macro F1 score of 0.87631. With corresponding scores of 0.874833, 0.86658, 0.84927 and 0.84237, HerWILL, One_by_zero, Dll5143, and MNLP all shown strong performance.

The diverse methodologies employed by teams such as teamiic, byteSizedLLM, Team_Strikers, InnovationEngineers, Zero_knowledge, Code_Conquerors, HERWILL, Shraddha, CUET-NLP__Big_O, CUET-823 ,One_by_zero, CUET-NLP_MP and CUET_NetworkSociety highlighted the significance of multimodal approaches, combining textual and visual features for effective classification. Many teams leveraged pre-trained language models such as IndicBERT, XLM-RoBERTa, MuRIL, and multilingual BERT for textual feature extraction, often fine-tuned for Tamil and Malayalam. To capture visual contexts on the image side, models such as ResNet, Vision Transformer (ViT), Swin Transformer, and EfficientNet were frequently employed. In order to successfully integrate textual and visual embeddings, fusion techniques included dynamic attention mechanisms as well as early and late fusion where used.

The team Dll5143 improved the performance of the model by ensemble methods like majority voting and concatenation of multimodal embeddings, while teams such as buteSizedLLM and CUET-823 used novel approaches such as transliteration-enhanced datasets, back-translation, and data augmentation for both text and images, tackled issues with low-resource languages and imbalanced datasets. models with attention mechanisms, BiLSTM, and GRU captured contextual subtleties, and most of the teams employed dropout regularization and adam optimizer along with other hyperparameters. Teams such as Shraddha and LexiLogic used focal loss addressed overfitting and class imbalance.

The teams such as Magma, LinguAIsts and DLRG_RR used text based model like SVM, mBERT and BERT for classifiction task along with optimizing parameters. The team semanticCuetSync leveraged the LLaMa 3.2 , a large vision model for classification.

6 Conclusion and Future Work

In this second shared task, we aimed to address the issue of identifying misogyny memes in Tamil and Malayalam languages. The results demonstrated that multimodal fusion-based techniques yield better results in both the Tamil and Malayalam language dataset when compared with other techniques. Among the teams that submitted the results, most of them extracted textual and visual features separately using their appropriate models such as XLM-RoBERTa, T5, IndicBERT, MURIL for textual features and ResNet18, ViT, CNN, EfficientNetB0 for visual features. The features are then has been combined using fusion techniques and fed to the classifier model. Even though, multimodal models performed well for this dataset, we plan to explore the bias like data, contextual and algorithmic in data. Models can be trained to understand the local cultural differences and sensitivity in data by annotating the data set with detailed context information. Fine grained multimodal analysis is needed for detecting misogyny memes because of the subtle change in tone, context and nuances present in the image and text present in the memes which cannot be detected on the surface level analysis. Future research in these areas can improve the detection of misogyny memes for the safer online environment.

Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2), supported in part of Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

Arthir, Pavithra J, G Manikandan, Lekhashree A4 Dhanyashree G, Bommineni Sahitya, Arivuchudar K, and Kalpana K. 2025. LinguAIsts@DravidianLangTech 2025: Misogyny Meme Detection using multimodel Approach. In *Proceedings of the Fifth Workshop on Speech,*

Vision, and Language Technologies for Dravidian Languages. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Dola Chakraborty, Shamima Afroz Mithi, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. One_by_zero@DravidianLangTech 2025: A Multimodal Approach for Misogyny Meme Detection in Malayalam Leveraging Visual and Textual Features. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of Shared Task on Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Shraddha Chauhan and Abhinav Kumar. 2025. MNLP@DravidianLangTech 2025: Transformer-based Multimodal Framework for Misogyny Meme Detection. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR. ISSN: 2640-3498.

Md. Sajid Alam Chowdhury, Mostak Mahmud Chowdhury, Anik Mahmud Shanto, Jidan Al Abrar, and Hasan Murad. 2025. Fired_from_NLP@DravidianLangTech 2025: A Multimodal Approach for Detecting Misogynistic Content in Tamil and Malayalam Memes. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Tommaso Caselli, Nicole Novielli, and Viviana Patti, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian : Proceedings of the Final Workshop 12-13 December 2018, Naples*, Collana dell'Associazione Italiana di Linguistica Computazionale, pages 59–66. Accademia University Press, Torino. Code: EVALITA Evaluation of NLP and Speech Tools for Italian : Proceedings of the Final Workshop 12-13 December 2018, Naples.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in Brief*, 44:108526.

Md. Refaj Hossan, Nazmus Sakib, Md. Alam Miah Jawad Hossain Hoque, and Mohammed Moshiul. 2025. CUET-NLP_Big_O@DravidianLangTech 2025: A Multimodal Fusion-based Approach for Identifying Misogyny Memes. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Jianzhao Huang, Hongzhan Lin, Liu Ziyan, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with LMM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2269–2293, Miami, Florida, USA. Association for Computational Linguistics.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. MemeFier: Dual-stage Modality Fusion for Image Meme Classification. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, pages 586–591, New York, NY, USA. Association for Computing Machinery.

Gitanjali Kumari, Kirtan Jain, and Asif Ekbal. 2024. M3Hop-CoT: Misogynous meme identification with multimodal multi-hop chain-of-thought. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22105–22138, Miami, Florida, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. ArXiv:1907.11692 [cs].

Niranjan Kumar M, Pranav Gupta, Billodal Roy, and Souvik Bhattacharyya. 2025. LexiLogic@DravidianLangTech 2025: Detecting Misogynistic Memes and Abusive Tamil and Malayalam Text Targeting Women on Social Media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sidharth Mahesh, Sonith D, Gauthamraj Gauthamraj, Kavya G, Asha Hegde, and H Shashirekha. 2024. MUCS@LT-EDI-2024: Exploring joint representation for memes classification. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 282–287, St. Julian's, Malta. Association for Computational Linguistics.

Arpita Mallik, Ratnajit Dhar, Udoy Das, Momtazul Arefin Labib, Samia Rahman, and Hasan Murad. 2025. CUET-823@DravidianLangTech 2025: Shared Task on Multimodal Misogyny Meme Detection in Tamil Language. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@DravidianLangTech 2025: Multimodal Misogyny Meme Detection in Low-Resource Dravidian Languages Using Transliteration-Aware XLM-RoBERTa, ResNet-50, and Attention-BiLSTM. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Md. Mohiuddin, Md Minhazul Kabir, Kawsar Ahmed, and Mohammedmoshiul Hoque. 2025. CUET-NLP_MP@DravidianLangTech 2025: A Transformer-Based Approach for Bridging Text and Vision in Misogyny Meme Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sarbajeet Pattanaik, Ashok Yadav, and Vrijendra Singh. 2025. Dll5143@DravidianLangTech 2025: Majority Voting-Based Framework for Misogyny Meme Detection in Tamil and Malayalam. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Parth Patwa, Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Findings of memotion 2: Sentiment and emotion analysis of memes. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Neelima Monjusha Preeti, , Trina Chakraborty, , Noor Mairukh Khan Arnob, , Saiyara Mahmud, , and Azmine Toushik and Wasi. 2025. HerWILL@DravidianLangTech 2025: Ensemble Approach for Misogyny Detection in Memes Using Pretrained Text and Vision Transformers. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR. ISSN: 2640-3498.

Pathange Omkareshwara Rao, Harish Vijay V, Ippatapu Venkata Srichandra, Neethu Mohan, and Sachin Kumar S. 2025. Code_Conquerors@DravidianLangTech 2025: Multimodal Misogyny Detection in Dravidian Languages Using Vision Transformer and BERT. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Giulia Rizzi, Alessandro Astorino, Paolo Rosso, and Elisabetta Fersini. 2024. Unraveling disagreement constituents in hateful speech. In *European Conference on Information Retrieval*, pages 21–29. Springer.

Khadiza Sultana Sayma, Farjana Alam Tofa, Md Osama Dey, and Ashim. 2025. CUET_Novice@DravidianLangTech 2025: A Multimodal Transformer-Based Approach for Detecting Misogynistic Memes in Malayalam Language. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Mohamed Arsath H, Ramya K, and Ragav R. 2025a. TEAM_STRIKERS@DravidianLangTech2025: Misogyny Meme Detection in Tamil Using Multimodal Deep Learning. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Pooja Sree M, Palanimurugan V, and Roshini Priya K. 2025b. InnovationEngineers@DravidianLangTech 2025: Enhanced CNN Models for Detecting Misogyny in Tamil Memes Using Image and Text Classification. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Harshita Sharma, Simran, Vajratiya Vajrobol, and Nitisha Aggarwal. 2025. teamiic@DravidianLangTech 2025: Transformer-Based Multimodal Feature Fusion for Misogynistic Meme Detection in Low-Resource Dravidian Language. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Smriti Singh, Amritha Haridasan, and Raymond Mooney. 2023. "female astronaut: Because sandwiches won't make themselves up there": Towards multimodal misogyny detection in memes. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 150–159, Toronto, Canada. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2023. TrollsWithOpinion: A taxonomy and dataset for predicting domain-specific opinion manipulation in troll memes. *Multimedia Tools and Applications*, 82(6):9137–9171.

Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu

Durairaj Thenmozhi[1], **Bharathi Raja Chakravarthi**[2], **Asha Hedge**[3], **Hosahalli Lakshmaiah Shashirekha**[3], **Rajeswari Natarajan**[4], **Sajeetha Thavareesan**[5], **Ratnasingam Sakuntharaj**[5] **Krishnakumari Kalyanasundaram**[6] **Charumathi Rajkumar**[7] **Poorvi Shetty**[8] **Harshitha S Kumar**[3]
[1]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India,
[2]School of Computer Science, University of Galway, Ireland,
[3]Mangalore University, Mangalore, India,
[4]SASTRA University, SRC campus, Kumbakonam, Tamil Nadu, India,
[5]Eastern University, Sri Lanka, [6]A.V.C. College of Engineering, Tamil Nadu, India
[7]The American College, Madurai, Tamil Nadu, India, [8]JSS College, Mysore, India.

Abstract

Sentiment analysis is an essential task for interpreting subjective opinions and emotions in textual data, with significant implications across commercial and societal applications. This paper provides an overview of the shared task on Sentiment Analysis in Tamil and Tulu, organized as part of DravidianLangTech@NAACL 2025. The task comprises two components: one addressing Tamil and the other focusing on Tulu, both designed as multi-class classification challenges, wherein the sentiment of a given text must be categorized as positive, negative, neutral and unknown. The dataset was diligently organized by aggregating user-generated content from social media platforms such as YouTube and Twitter, ensuring linguistic diversity and real-world applicability. Participants applied a variety of computational approaches, ranging from classical machine learning algorithms such as Traditional Machine Learning Models, Deep Learning Models, Pre-trained Language Models and other Feature Representation Techniques to tackle the challenges posed by linguistic code-mixing, orthographic variations, and resource scarcity in these low resource languages.

1 Introduction

Sentiment Analysis (SA), a computational approach to deciphering human opinions and emotions through written language, has become increasingly crucial across various domains such as social media monitoring, market research, and customer feedback analysis (Wankhade et al., 2022). The digital communication landscape has witnessed a growing linguistic phenomenon known as code-mixing, which is particularly prevalent in multilingual societies (Alam et al., 2024). Code-mixing represents the intricate practice of interweaving multiple languages within a single communicative context, a trend that has garnered significant attention in Natural Language Processing, especially in linguistically diverse regions like India (Sampath and Supriya, 2024). An enhanced sentiment dictionary that incorporates both labeled and unlabeled data from source and target domains significantly improves sentiment classification in multi-domain contexts (Sivasankar et al., 2021). This approach is not only applicable to domain adaptation within the same language but can also be extended to sentiment classification across different languages. By leveraging such cross-lingual adaptation, models can better handle language-specific nuances, improving sentiment analysis in low-resource languages like Tamil and Tulu, as well as facilitating sentiment classification between languages with distinct linguistic features. Systems trained on monolingual data face challenges with code-mixed text because of the intricate nature of code-switching across different linguistic levels. (Ponnusamy et al., 2023). The complexity of SA escalates when confronting code-mixed text, as traditional analysis methods struggle with the nuanced linguistic variations introduced by script and language mixing (Perera and Caldera, 2024). Users frequently leverage Latin script and common English words, creating hybrid textual landscapes that challenge conventional sentiment extraction techniques (Hegde et al., 2022). The intricate nature of code-mixing further compounds sentiment analysis challenges, with individuals often switching between scripts and languages in unpredictable ways (Chakravarthi et al., 2021; Sambath Kumar et al., 2024). These linguistic complexities highlight the critical need for more advanced analytical

approaches. Addressing these code-mixed linguistic complexities is important to capture modern communication's rich, dynamic nature accurately.

Granted classical language status by the Indian government in 2004, Tamil boasting a literary heritage that spans over two millennia and is one of the world's most enduring classical languages (Abirami et al., 2024). Beyond its status as the official language of Tamil Nadu and Puducherry, the language has transcended geographical boundaries, finding vibrant expression in diverse global communities including Malaysia, Mauritius, Fiji, and South Africa (Rajalakshmi et al., 2023).

Tulu, a member of the Dravidian language family, boasts over three million speakers known as Tuluvas, primarily concentrated in Karnataka's Dakshina Kannada and Udupi districts, with additional communities extending to Mumbai and Gulf countries (Hegde et al., 2022). The language has carved out a significant digital footprint, with active engagement across social media platforms and a thriving film industry that further amplifies its cultural relevance (Narayanan and Aepli, 2024).

This shared task presents a new corpus in the Tamil and Tulu languages. We used comments and posts of Movie reviews from Youtube for this shared task of Sentiment Analysis.

2 Task Description

The goal of this shared task[1] is to identify the sentiment polarity of the code-mixed dataset of comments or posts in Tamil-English and Tulu-English collected from social media. The comment or post may contain more than one sentence but the average sentence length of the corpora is one. Each comment or post is annotated with sentiment polarity at the comment or post level. These code-mixed datasets consist of posts and comments collected from YouTube comments. Our proposal aims to encourage research that will reveal how sentiment is expressed in code-mixed scenarios on social media. For every comment in Tamil and Tulu, the objective is to classify it into positive, negative, neutral, or mixed emotions.

3 Dataset Description

Recent advancements in natural language processing (NLP), particularly transformer-based models and multilingual embeddings, have further accelerated research in sentiment analysis. Additionally, the integration of large language models (LLMs) and zero-shot learning techniques has improved sentiment classification accuracy for underrepresented languages, enabling better contextual understanding and real-time analysis. For the analysis of sentiment in YouTube comments, two meticulously curated datasets—i) Tamil-English and ii) Tulu-English—have been introduced to support computational linguistics research. These resources act as essential linguistic benchmarks, aiding in the exploration of hybrid-language processing. By offering diverse and naturally occurring text samples, they assist scholars and industry professionals in enhancing machine learning models tailored for multilingual and phonetically transcribed content. Additionally, these datasets play a pivotal role in refining Artificial Intelligence systems to better interpret the emotional tone and contextual intricacies of underrepresented languages in digital discourse. The dataset is prepared in two languages such as Tamil and Tulu as listed in Table 1.

3.1 Tamil Data

We gathered comments and posts from YouTube related to various Tamil films, encompassing discussions, reviews, and audience opinions. This data includes user perspectives on different aspects such as storyline, performances, music, and overall cinematic experience. A sample set of comments are listed in the Table 2.

3.2 Tulu Data

We collected comments and posts from YouTube about various Tulu films, covering discussions, reviews, and audience opinions. The data reflects user perspectives on elements such as storyline, performances, music, and overall cinematic appeal. A sample set of comments are listed in the Table 3.

4 Methodologies used in the Submission

Team Hermes fine-tuned the pre-trained multilingual transformer model, cardiffnlp/twitter-xlm-roberta-base-sentiment from Hugging Face. A PyTorch-based data pipeline with a custom dataset class and DataLoaders was used to handle batched input. The model employed AdamW optimization, and early stopping based on validation F1 score.

Team byteSizedLLM (Manukonda and Kodali, 2025) used hybrid approach combined a fine-tuned XLM-RoBERTa base model with a customized attention BiLSTM network to leverage contextualized embeddings and sequential modeling. The

[1]https://codalab.lisn.upsaclay.fr/competitions/20893

Label	Train Set		Development Set		Test Set	
	Tamil	Tulu	Tamil	Tulu	Tamil	Tulu
Positive	18,145	3,769	2,272	470	1,983	453
Negative	4,151	843	480	118	458	88
Neutral	5,164	3,175	619	368	593	343
Mixed	3,662	1,114	472	143	425	120

Table 1: Distribution of data in Train, Development, and Test sets for Tamil and Tulu languages

S.No	Text	Label
1	Therikaaa vidalamaanu kealvi yellam keadayadhu... Iranginaale theri dhaaaa	Positive
2	Romba naalaki aprama suriya annana ipdi pakuravanga mattum solluga	Unknown_state
3	Aiooo..samy mudilada..yenda 2D ku inoru flop conform	Negative
4	The word vera level thalaiva unaku vayase agadha paaaa	Mixed_feelings

Table 2: Sample set of Tamil Comments with Labels

XLM-RoBERTa model was fine-tuned using MLM on a small portion of the AI4Bharat dataset, enriched with fully and partially transliterated text to improve multilingual and transliteration handling. Attention and BiLSTM layers were used to enhance sequential dependency capture.

For the Tamil dataset, XNet, Naive Bayes, and Logistic Regression were employed by Team RMK-Mavericks to predict sentiment, leveraging their ability to capture diverse text patterns and nuances. For the Tulu dataset, they used SVM, Random Forest, and Logistic Regression. TF-IDF vectorization transformed text into numerical features, with hyperparameter tuning to optimize results.

Team codecrackers (P et al., 2025) system employs three models—Naive Bayes, SVM, and an LSTM-based deep learning model—to address sentiment analysis in Tamil code-mixed text. Naive Bayes and SVM leverage TF-IDF vectorization and traditional machine learning techniques for simpler patterns, while the LSTM-based model combines word-level Word2Vec embeddings and character-level features to capture complex syntactic and semantic nuances.

(Sreeja and Bharathi, 2025) fine-tuned a pre-trained transformer model, distilroberta-base, for multilingual sentiment analysis. Their program preprocesses data by cleaning text, mapping labels to numeric values, and tokenizing inputs, while addressing class imbalances by incorporating calculated class weights into the loss function. Optimized training techniques like gradient accumulation and mixed precision enhanced efficiency.

Team Code Conquerors employed data preprocessing that involved addressing class imbalance using the class-weight method, and resolving out-of-vocabulary issues by developing a vocabulary. The hybrid model begins with an embedding layer, followed by a Conv1D layer and a max pooling layer. A BiLSTM layer was used, with fully connected dense layers and a softmax layer for finding the sentiments.

Team ET2025 used mBERT model (Adyanthaya, 2025) fine-tuned for sentiment classification in Tamil and Tulu. Preprocessed datasets were tokenized and split into training and evaluation sets, with the Hugging Face Trainer API used for training and evaluation.

Team CIC used Feature extraction with Logistic Regression which involved identifying the top bigrams and trigrams based on frequency to capture class similarities effectively. These enabled the model to distinguish between classes more accurately.

Team SKV Trio combined TF-IDF and BERT embeddings. TF-IDF embeddings were reduced to 512 dimensions via an RBF Sampler for efficiency, while the bert-base-uncased model generated con-

S.No	Text	Label
1	Edde msg koryar prasamsha	Positive
2	Enchi pankda comedy	Negative
3	Kas ejjande boys yerla hotel popujer. Ponnulu mathra	Mixed
4	Padhyana ganapathi bhagavathike	Neutral
5	Well done Keep It up!!!	Not Tulu

Table 3: Sample set of Tulu Comments with Labels

textual embeddings by averaging token representations. Random Forest classifier was trained on the merged features. Model generalizability was verified through 5-fold cross-validation.

Pre-processing techniques such as tokenization, special character removal, and stopword filtering were applied by Team YenLP_CS. TF-IDF was used for feature extraction (Adyanthaya, 2025), followed by training an ensemble of Random Forest (fine-tuned with GridSearchCV) and SVM (Shanmugavadivel et al., 2025). Word2Vec embeddings were generated, supporting LSTM and BiLSTM deep learning models. Finally, the multilingual transformer model mBERT was fine-tuned on the task data.

Team KECTechTitans used three models - KNN, SVM, and Decision Tree on data vectorized by TF-IDF. KNN was selected for its simplicity in proximity-based classification, SVM for its effectiveness in high-dimensional data, and Decision Tree for its interpretability. Model performance compared to determine the most effective approach.

Team Dynamic_Crew cleaned text data and normalized it to remove noise, such as special characters and stop words, while addressing language-specific nuances. For feature extraction, Count Vectorizer and TF-IDF Vectorizer were used. Classifiers like Decision Tree, Random Forest, and KNN were employed by this team.

Team Team_Mavericks used TF-IDF Vectorizer to capture term importance with unigrams and bigrams. Data preprocessing included combining training and validation datasets, splitting for evaluation, and vectorizing the text. The Random Forest model was used, optimized using GridSearchCV for hyperparameter tuning.

Team JustATalentedTeam (Ponsubash Raj R, 2025) began with transliterating texts to the English script when necessary. Two methodologies were used: 1) a Logistic Regression model with TF-IDF Vectorizer using a character-level analyzer, and 2) a combined approach involving tokenization, FastText embeddings, and a deep learning model for sentiment classification of code-mixed text.

Team lemlem used google-bert/bert-base-multilingual-uncased model, with a classification head consisting of a dense layer and softmax activation to predict sentiment categories. By fine-tuning the model, the system aimed to subtle syntactic and semantic nuances, performing sentiment analysis even with limited annotated data.

Team MysticCIOL's approach involved using custom pre-trained models, each specifically trained on general Tamil and Tulu data. A Multi-Layer Perceptron was applied on top of the pre-trained embeddings to fine-tune them for sentiment classification. The fine-tuned models were then used to generate predictions on the test data.

In Team Cognitext's approach, the text data was preprocessed by converting it to lowercase, removing URLs, mentions, hashtags, and special characters. TF-IDF vectorization was applied to extract feature through unigrams and bigrams. A Logistic Regression classifier was then trained on it. The model's performance was evaluated using a validation set.

Team TensorTalk (Anishka and J, 2025) used a combination of SVM, Logistic Regression, and Random Forest classifiers. The preprocessing pipeline involved cleaning the text, removing stopwords, performing lemmatization, and applying TF-IDF vectorization. To address the class imbalance in the datasets, we used the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic samples for underrepresented classes.

Team SSNTrio (J et al., 2025) explored the use of Multilingual BERT and language-specific Tamil

BERT models. To address the class imbalance, random upsampling was applied. The text data was tokenized using the appropriate tokenizer for each model. The models were fine-tuned using the training set, with optimized hyperparamters.

Team Anna-CIOL used custom pre-trained models specifically fine-tuned for Tamil and Tulu to extract embeddings. They employed a Multi-Layer Perceptron to fine-tune these embeddings for sentiment classification. Once the models were fine-tuned, they used them to generate predictions.

Team lowes fine-tuned a language-specific BERT model pre-trained by l3cube-pune using the provided datasets, implementing class weighting to handle imbalanced labels and optimizing training parameters. The model handled class imbalance through oversampling and by using a weighted model to compute loss during training.

The preprocessing techniques, feature extractions, and classifiers used by the participating teams are summarized below.

4.1 Preprocessing

Teams have used preprocessing techniques namely stop word removal, removal of hashtags & URLS and lemmatization in their approaches. They have used random upsampling and SMOTE methods for handling data imbalance problems.

4.2 Feature Extraction

Submitting teams used unigrams, bigrams, trigrams and Character-level features, and vectorized the text with TF-IDF, Word2Vec, fasttext and BERT embeddings.

4.3 Classifiers

Participants used traditional classifiers namely, SVM, logistic regression, multilayer perceptron, random forest (Gowda, 2025), decision tree, KNN and Naive Bayes approaches for finding the sentiments. Deep learning frameworks namely LSTM and BILSTM are used by the participants (Srichandra et al., 2025; Rajalakshmi et al., 2023). Transformer models namely distilRoberta, Multilingual BERT and XLM-Roberta are used by the participants among which XLM-Roberta performs better when compared to other approaches (Krasitskii et al., 2025) (G et al., 2025). Further, language specific BERT pretrained by l3cube-pune performs better for Tulu language with 0.5938 as F1-score. A pretrained model finetuned on AI4Bharart dataset performs better for Tamil language with an F1-score of 0.5036.

5 Results

The participating teams submitted 2 to 3 runs to the both Tamil and Tulu tasks. 21 teams participated in Tamil task and 20 teams participated in Tulu task. Their submissions were evaluated and ranked based on macro F1-score. Scores are tabulated in Tables 4 and 5 for Tamil and Tulu sub tasks respectively.

Team Name	Macro F1 Score	Rank
byteSizedLLM	0.5036	1
ET2025	0.4986	2
Hermes	0.4957	3
JustATalentedTeam	0.4919	4
Lemlem	0.4709	5
SSNTrio	0.4461	6
CIC	0.4409	7
codecrackers	0.4389	8
KECTechTitans	0.4386	9
RMKMavericks	0.4354	10
MysticCIOL	0.4299	11
KEC-Elite-Analysts	0.4131	12
YenLP_CS	0.4117	13
Team_Mavericks	0.4011	14
Dynamic_crew	0.3852	15
lowes	0.3834	16
SSN_IT_SENTI	0.3799	17
CodeConquerors	0.3357	18
Anna-CIOL	0.335	19
Cognitext	0.2867	20
TensorTalk	0.2427	21

Table 4: Rank list for Tamil sentiment analysis.

6 Conclusion

There are 21 teams participated in the Tamil subtask and 20 teams participated in the Tulu subtask. Participants used preprocessing techniqes like lemmatization, removal of stop word, URLs and hash tags in their approaches. Teams who have employed traditional classifiers used n-gram features with TF-IDF scores and static embeddings namely Word2Vec and FastText for vectorization. Most of the teams used transformer models namely multilingual BERT, distilRoberta and XLM-Roberta, Pretrained models finetuned on AI4Bharat dataset are used by the teams. Team 'byteSizedLLM' (Manukonda and Kodali, 2025) who used language

Team Name	Macro F1 Score	Rank
lowes	0.5938	1
ET2025	0.5882	2
Hermes	0.5801	3
JustATalentedTeam	0.5617	4
SSNTrio	0.5609	5
Lemlem	0.5583	6
YenLP_CS	0.5511	7
codecrackers	0.5425	8
RMKMavericks	0.5318	9
TensorTalk	0.5269	10
Team_Mavericks	0.4683	11
CIC	0.4509	12
CodeConquerors	0.4357	13
SSN_IT_SENTI	0.3904	14
Anna-CIOL	0.3863	15
SKV-trio	0.3767	16
Dynamic_crew	0.3750	17
KECTechTitans	0.3197	18
MysticCIOL	0.1546	19
Cognitext	0.1491	20

Table 5: Rank list for Tulu sentiment analysis.

BERT pretrained on AI4Bharat dataset secure first position in the Tamil subtask, and the team "lowes" who used pretrained models created by l3cube-pune secured first position in the Tulu subtask.

Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2), supported in part of Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

A M Abirami, Wei Qi Leong, Hamsawardhini Rengarajan, D Anitha, R Suganya, Himanshu Singh, Kengatharaiyer Sarveswaran, William Chandra Tjhi, and Rajiv Ratn Shah. 2024. Aalamaram: A large-scale linguistically annotated treebank for the Tamil language. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pages 73–83, Torino, Italia. ELRA and ICCL.

Raksha Adyanthaya. 2025. Sentiment analysis on code-mixed tamil and tulu data using machine learning and deep learning models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee, Md Shahnewaz Siddique, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2024. BnSentMix: A diverse Bengali-English code-mixed dataset for sentiment analysis.

K Anishka and Anne Jacika J. 2025. Sentiment analysis in tamil and tulu-dravidianlangtech@naacl2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Anand Kumar Madasamy, Sajeetha Thavareesan, Bhavukam Premjith, K R Sreelakshmi, Subalalitha Chinnaudayar Navaneethakrishnan, John, Patrick McCrae, and Thomas Mandl. 2021. Overview of the hasoc-dravidiancodemix shared task on offensive language detection in tamil and malayalam. In *In Proceedings of the Forum for Information Retrieval and Evaluations*.

Jyothish Lal G, Premjith B, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Ratnavel Rajalakshmi. 2025. Overview of the shared task on multimodal hate speech detection in dravidian languages: Dravidianlangtech@naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Anusha M D Gowda. 2025. Bridging linguistic complexity: Sentiment analysis of tamil code-mixed text using meta-model. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Bhuvana J, Mirnalinee T T, Diya Seshan, Rohan R, and Avaneesh Koushik. 2025. Ssntrio@dravidianlangtech 2025: Sentiment analysis in dravidian languages using multilingual bert. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Mikhail Krasitskii, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2025. Multilingual sentiment analysis: Understanding tamil-english code-mixing with transformer models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language*

Technologies for Dravidian Languages. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. Sentiment analysis in tamil using transliteration-aware xlm-roberta and attention-bilstm. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Manu Narayanan and Noëmi Aepli. 2024. A Tulu resource for machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1756–1767, Torino, Italia. ELRA and ICCL.

Lalith Kishore V P, Dr. G Manikandan, Mohan Raj M A, Keerthi Vasan A, and Aravindh M. 2025. odecrackers@dravidianlangtech 2025: Sentiment classification in tamil and tulu code-mixed social media text using machine learning. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Perera and Caldera. 2024. Sentiment analysis of code-mixed text: A comprehensive review. *Int. J. Comput. Sci. Netw. Secur.*, 24(11):73–84.

Kishore Kumar Ponnusamy, Charmathi Rajkumar, Prasanna Kumar Kumaresan, Elizabeth Sherly, and Ruba Priyadharshini. 2023. Vel@ dravidianlangtech: Sentiment analysis of tamil and tulu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 211–216.

Bharathi B Ponsubash Raj R, Paruvatha Priya B. 2025. Justatalentedteam@dravidianlangtech 2025: A study of ml and dl approaches for sentiment analysis in code-mixed tamil and tulu texts. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Faerie Mattins, Pavitra Vasudevan, and Anand Kumar. 2023. HOTTEST: Hate and offensive content identification in tamil using transformers and enhanced STemming. *Comput. Speech Lang.*, 78(101464):101464.

Lavanya Sambath Kumar, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 2024. Overview of second shared task on sentiment analysis in code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 62–70, St. Julian's, Malta. Association for Computational Linguistics.

Koyyalagunta Krishna Sampath and M Supriya. 2024. Transformer based sentiment analysis on code mixed data. *Procedia Comput. Sci.*, 233:682–691.

Kogilavani Shanmugavadivel, Malliga Subramanian, Sanjai R, Mohammed Sameer, and Motheeswaran K. 2025. Beyond_tech@dravidianlangtech 2025: Political multiclass sentiment analysis using machine learning and neural network. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

E Sivasankar, Kalyanasundaram Krishnakumari, and P Balasubramanian. 2021. An enhanced sentiment dictionary for domain adaptation with multi-domain dataset in tamil language (esd-da). *Soft Computing*, 25:3697–3711.

K Sreeja and B Bharathi. 2025. Multimodal hate speech detection in dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ippatapu Venkata Srichandra, Harish Vijay V, Pathange Omkareshwara Rao, and Premjith B. 2025. Deep learning approach for sentiment analysis in tamil and tulu – dravidianlangtech@naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.*, 55(7):5731–5780.

Overview on Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments: DravidianLangTech@NAACL 2025

Bharathi Raja Chakravarthi[1], Saranya Rajiakodi[2], Elizabeth Sherly[3], Thenmozhi Durairaj[4], Sathiyaraj Thangasamy[5], Ratnasingam Sakuntharaj[6], Prasanna Kumar Kumaresan[7], Kishore Kumar Ponnusamy[3], Arunaggiri Pandian Karunanidhi[8], Rohan R[4],

[1]University of Galway, Ireland,
[2]Central University of Tamil Nadu, India, [3]Digital University Kerala, Kerala, India,
[4]Sri Sivasubramaniya Nadar College of Engineering, Chennai,
[5]Sri Krishna Adithya College of Arts and Science, India, [6]Eastern University, Sri Lanka.
[7]Data Science Institute, University of Galway, Ireland, [8]Micron Technology, United states.

Abstract

Political multiclass detection is the task of identifying the predefined seven political classes. In this paper, we report an overview of the findings on the "Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments" shared task conducted at the workshop on DravidianLangTech@NAACL 2025. The participants were provided with annotated Twitter comments, which are split into training, development, and unlabelled test datasets. A total of 139 participants registered for this shared task, and 25 teams finally submitted their results. The performance of the submitted systems was evaluated and ranked in terms of the macro-F1 score.

1 Introduction

Online platforms are becoming the key platforms for the public conversation and the distribution of political news due to the quick development of digital and social media (Hermida et al., 2012; Kümpel et al., 2015; Tumasjan et al., 2010). Users may voice their thoughts, participate in conversations, and organize political movements with a reach and involvement previously unavailable on platforms like X (formerly Twitter) (Mustafaraj and Metaxas, 2011; Velasquez, 2012). Over the past decade, social media has fueled conversations on a wide range of divisive political topics, including climate change, gun control, abortion rights, income inequality, the death penalty, taxation policies, and LGBTQ+ rights (Rainie et al., 2012; Zhuravskaya et al., 2020). In addition to encouraging democratic participation and a range of ideas, these conversations often serve to magnify social prejudices, frequently reinforcing divisive opinions and political divisions (Blair, 2002; Devine, 1989).

As online political discourse expands, Natural language processing (NLP) models are increasingly being used to analyze public sentiment and opinion trends. However, many of these models are trained on vast datasets gathered from online sources, which inherently reflect existing societal biases. Political sentiment analysis is not solely a technological challenge but also involves issues of fairness and the ethical application of AI. (Blodgett et al., 2020; Kumar et al., 2022; Field et al., 2021). Numerous research studies have emphasized the dangers of bias in NLP models, such as incorrect sentiment categorization, unintentional reinforcement of ideological viewpoints, and distortion of minority voices (Nangia et al., 2020; Sun et al., 2019). Moreover, the subjective character of political state-of-mind labeling and differences in annotator viewpoints make attempts to create objective models much more challenging (Feng et al., 2023; Sap et al., 2019).

Sentiment analysis has advanced, but political expression poses special difficulties that need advanced strategies. Political conversations frequently contain sarcasm, coded language, and shifting rhetorical methods that are challenging for standard models to accurately interpret, unlike generic sentiment classification tasks where text is simply categorized as positive, negative, or neutral (Demszky et al., 2019). Furthermore, the framing of language is influenced by biases in political reporting and media coverage, making it significantly harder to train objective sentiment analysis models. (Joseph and Morgan, 2020).

This work presents a summary of the Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments shared task, which intends to improve multilingual and low-resource sentiment analysis research in order to overcome these issues. This work offers a chance to investigate the shortcomings of existing AI techniques for expressing sentiment in political situations by concentrating on Tamil, a linguistically rich language. The objective

is to compare different strategies, find limitations in current techniques, and encourage improvements in the categorization of political perspective for under-resourced languages. We collected a dataset containing Tamil comments from X(Twitter) and then annotated the dataset for seven predefined classes. Then, we split it into training, development, and test sets for this task.

2 Related work

Several studies have explored sentiment analysis in Tamil, particularly focusing on social media platforms like Twitter. For instance, the study Anbukkarasi and Varadhaganapathy (2020) employed deep learning algorithms such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) to analyze Tamil tweets, achieving notable accuracy and F1-scores.

Another study, Thavareesan and Mahesan (2019) investigated various machine learning approaches for sentiment classification in Tamil texts, contributing to the understanding of effective methods for Tamil sentiment analysis. Additionally, Shanmugavadivel et al. (2022) addressed the challenges of analyzing sentiments in code-mixed Tamil texts, which are common in social media contexts. This study utilized machine learning techniques to classify sentiments in such code-mixed data.

Furthermore, Mahata et al. (2020) explored sentiment classification in code-mixed Tamil-English tweets using a Bi-Directional Recurrent Neural Network (RNN) approach, highlighting the complexities and solutions in handling mixed-language data. In addition to these studies, (Anish and Sumathy) proposed an SVM-based approach to analyze sentiments in Tamil political reviews, while (Devasena et al., 2022) demonstrated how sentiment analysis could be applied to predict election results based on Twitter data.

Sharmista and Ramaswami (2020), examined Tamil sentiment classification in the context of product reviews, showcasing its relevance in different domains. Lastly, (Shanmugavadivel et al., 2022) explored embedding representations for code-mixed Tamil text, addressing the challenges posed by multilingual and informal social media content.

These studies collectively contribute to the advancement of sentiment analysis methodologies for Tamil, particularly in the context of social media data. However, limited research exists on political multiclass sentiment analysis in Tamil, which involves classifying sentiments into multiple nuanced categories beyond the traditional positive, negative, and neutral classes. Our work aims to bridge this gap by introducing a detailed classification scheme tailored to Tamil political discourse.

3 Task Description

The primary goal of this task is to detect the political categories in the comments collected from X (Twitter). The participants were provided with training, development, and test datasets. The dataset is tagged using 7 classes namely, Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral and None of the above. Further information on the task is available in the Codalab site[1].

3.1 Datasets

The dataset containing Tamil text is the social media comments collected from X(Twitter). The diverse political sentiments captured in the dataset aim to reflect real-world nuances, making it well-suited for the multiclass sentiment analysis task. The dataset was divided into training, development, and testing sets. Training and validation sets are provided with class labels, and test sets are provided as unlabeled ones for evaluation. The data distribution and class distribution of training, validation, and test sets are given in Table 1

Table 1: Data Distribution

Class	Train	Development	Test	Total
Substantiated	412	52	51	515
Sarcastic	790	115	106	1,011
Opinionated	1,361	153	171	1,685
Positive	575	69	75	719
Negative	406	51	46	503
Neutral	637	84	70	791
None of the above	171	20	25	216
Tamil	4,352	544	544	5,440

4 Methodology

Totally 25 teams have actively participated in this shared task to detect the political comments in tamil. The participants have explored a variety of methodologies to classify the given comment as predefined political classes

Synapse team (KP et al., 2025) focused on preprocessing and fine-tuning to address class imbal-

[1] https://codalab.lisn.upsaclay.fr/competitions/20702

ance and optimize performance. During preprocessing, they converted emojis to text and expanded the top 160 most repeated hashtags to their full forms for better semantic understanding. For the model, they have finetuned IndicBERTv2-MLM-Back-TLM encoder based LLM model which was trained on IndicCorp v2 and Samanantar datasets, and an additional task of Translation. The fine-tuning was performed using the AutoModelForSequenceClassification architecture, incorporating class weights to rectify the class imbalance effectively. The team utilized only the train dataset for this fine-tuning process.

KCLR team (Mia et al., 2025) adopted a transformer-based deep learning architecture enhanced with multi-faceted embedding techniques. This approach combines three distinct feature extraction methods: attention-weighted representations, and CLS token embeddings from the transformer outputs. These features are concatenated to create comprehensive sentence representations before being processed through a fully connected classification layer. To address data imbalance challenges, the team implemented oversampling of minority classes using scikit-learn's resample function, ensuring robust and balanced training across all categories. This integrated approach, combining advanced feature engineering with balanced training data, enables effective multi-class classification while maintaining model robustness.

byteSizedLLM team implemented an advanced hybrid methodology combining a customized attention BiLSTM network with an XLM-RoBERTa base model, which had already been fine-tuned on the AI4Bharat dataset using Masked Language Modeling (MLM). The AI4Bharat dataset included fully and partially transliterated text, with 20–70 percentage of words randomly transliterated, enhancing transliteration-based diversity. This approach allows it to learn robust cross-lingual representations and adapt to varied transliteration patterns. The team further fine-tuned this pre-trained model and integrated BiLSTM and attention layers to capture sequential dependencies, making the model highly effective for multilingual and transliteration-heavy tasks.

Eureka-CIOL team (Eram et al., 2025) began by analyzing the dataset and identified that it consists of Tamil text with six distinct sentiment classes. Their best-performing model utilized a multilingual custom model pre-trained on general Twitter sentiment data, which allows for handling the diverse nature of social media content. To adapt the model for the specific task of sentiment classification, they applied a Multi-Layer Perceptron (MLP) on top of this pre-trained model, enabling fine-tuning. This approach leveraged the multilingual capabilities of the model and the domain-specific knowledge from general sentiment data. Finally the fine-tuned model was used to generate the predictions on the test data.

Wictory team (K et al., 2025) employed specific preprocessing techniques to prepare the data, including demojifying the text and removing unwanted characters. For their model, they converted word embeddings for generated LaBSE (Language-agnostic BERT Sentence Embedding), which were then passed into a Support Vector Machine (SVM) for classification.

MNLP team implemented the Deep Learning based model which was fine-tuned for classification. Their model achieved a 0.3026 macro F1-score and ranked 6th in the shared task.

Nova Spark developed a text classification pipeline for Tamil and English, involving text normalization, tokenization, and TF-IDF vectorization. To handle class imbalance, Borderline-SMOTE, SMOTEENN, and ADASYN were used. An optimized Support Vector Classifier (SVC) was trained using GridSearchCV for the best macro F1-score. Performance was evaluated with a classification report, and final predictions were saved as a CSV for submission.

Team_Catalysts (Shanmugavadivel et al., 2025a) implemented a robust Tamil text classification pipeline, including Unicode normalization, tokenization with Stanza, and standardization of spoken variants. Class imbalance was addressed through upsampling, followed by TF-IDF vectorization. A Random Forest Classifier was trained using stratified splitting and evaluated with accuracy and classification reports, ensuring effective sentiment analysis.

Lowes team began by preprocessing the dataset to prepare it for analysis. They then fine-tuned a BERT-based model specifically for the task. Their model achieved a 0.2908 macro F1-score and ranked 9th in the shared task.

Abhay43 team applied simple preprocessing to the dataset. They then extracted embeddings using the DeBERTa v3 model, which were subsequently fed into a two-layered LSTM model. They achieve a macro F1-score of 0.2904 and ranked tenth

GS Team explored several machine learning ap-

proaches including Logistic regression, random forest classifier, support vector machine, and XGBBoost classifier with TFIDF vectorization techniques for feature extraction techniques. Among these models, the XGBoost model outperformed the other models. Similarly, **JAS** team employed Logistic Regression as a primary approach for this classification task.

SentiTamil team utilized classical machine learning approaches, specifically support vector machine (SVM), with TFIDF vectorizer, limiting the number of features to 5,000 for efficiency. They also tried to fine-tune the tamil-llama-7b model, however the predicted value is not similar as the gold label of the training dataset.

CrewX team leveraged IndicBERT, a multilingual language model tailor for Indian languages, as the backbone for political multiclass sentiment analysis of Tamil Twitter comments. The dataset was preprocessed to handle challenges such as code-mixing, transliteration, and noise typical in social media text. Tokenization was performed using IndicBERT's tokenizer to preserve linguistic nuances. The team fine-tuned the pre-trained IndicBERT model on the DravidianLangTech dataset, utilizing a classification head with softmax activation to predict sentiment classes. To enhance performance, they experimented with techniques like data augmentation, stratified sampling, and weighted loss to address class imbalance. The model was trained using cross-entropy loss and optimized with AdamW, while employing early stopping to prevent overfitting. Evaluation metrics, including accuracy, F1-score, and precision-recall, were used to assess the model's effectiveness. This approach leverages IndicBERT's contextual understanding to address the intricacies of Tamil sentiment analysis in a political context.

AnalysisArchitects team (Jayaraman et al., 2025) implemented a diverse methodology by employing Naive Bayes, SVM, and LSTM models for the task of multiclass sentiment analysis. For the Naive Bayes approach, the team preprocessed the text, transformed it using CountVectorizer, and trained the model for multiclass sentiment analysis. Predictions were then generated on a test dataset, and the results were saved as a CSV file. The SVM model utilized TF-IDF features for text representation. After preprocessing the text, the team trained an SVM classifier and evaluated its performance on a test dataset. Predictions for the separate test set were also saved as a CSV file. This method tokenizes and pads Tamil text sequences, then trains an LSTM model for sentiment analysis. The model uses an embedding layer, LSTM for sequence learning, and a softmax output for classification. Input dimensions are adjusted, and sequence values are clipped to stay within valid range

Beyond_tech team (Shanmugavadivel et al., 2025b) utilized a combination of natural language processing techniques and pattern recognition to extract relevant information and generate appropriate responses. The methodology involved analyzing the task description and context, followed by segmenting the input into smaller, manageable parts. Each segment was processed to identify key concepts and relationships, facilitating the formulation of precise and coherent outputs. To ensure continuous improvement, the team applied an iterative feedback loop for refinement and alignment with task requirements. This approach allowed for efficient handling of complex queries, maintaining accuracy and clarity in response generation.

CUET_Novice team (Barua et al., 2025) utilized multiple deep learning architectures for their methodology. In the first approach (run1), they utilized a model with stacked Bidirectional GRU (BiGRU) layers, followed by normalization and a feedforward neural network for classification. In the second approach (run2), they utilized a model with multiple Bidirectional LSTM (BiLSTM) layers, similarly they applied normalization and a feedforward neural network. For the third approach (run3), they employed a transformer-based model, leveraging its advanced contextual understanding capabilities. This diverse experimentation with GRUs, LSTMs, and transformers allowed the team to explore various architectures for optimal

KSK team (M et al., 2025) implemented an incremental and continual learning for political multiclass sentiment analysis of Tamil tweets focusing on adapting models to new data while retaining prior knowledge. Algorithms like Stochastic Gradient Descent (SGD) and Online Naive Bayes dynamically update parameters for evolving sentiments. The team also utilized Incremental SVMs and Hoeffding Trees, enabling efficient updates without retraining on the entire dataset. Pretrained models like multilingual BERT are fine-tuned continually to adapt to new linguistic patterns while avoiding catastrophic forgetting. Online ensemble methods further enhance robustness, making them suitable for evolving Twitter data streams.

QuanNguyen team utilized the BERT multilin-

gual base model (cased) to perform multiclass sentiment analysis on Tamil X (Twitter) comments. The data preprocessing involved identifying and categorizing hashtags and icons uniquely associated with each sentiment class while removing special characters and irrelevant symbols for cleaner input. The multilingual BERT model, well-suited for handling multiple languages including Tamil, was fine-tuned on the preprocessed dataset to capture contextual and semantic patterns in sentiment. While BERT formed the core of the system, the team noted the potential for exploring other deep learning models to further enhance performance.

Team_Luminaries_0227 team began by preprocessing the dataset, including cleaning text data. They utilized the TF-IDF vectorizer to convert the textual data into numerical representations. To address class imbalance in the dataset, They applied the SMOTE (Synthetic Minority Oversampling Technique) algorithm, ensuring balanced class distributions. For classification, a Random Forest classifier was trained, with performance evaluated using metrics such as precision, recall, and F1-score. The trained models were saved for later use, and predictions were generated on the test dataset, ensuring the methodology aligns with the objective of the task.

VKG_VELLORE INSTITUTE OF TECHNOLOGY team utilized classification pipeline by extracting features from a pre-trained Indic-BERT language model, and then DBOW and TF-IDF methods were applied followed by CatBoost classifier for text classification. For better performance, they performed preprocessing steps like removing special characters and converting text to lowercase. After tokenizing the text using the BERT tokenizer, Indic-BERT embeddings were created, transforming the input text into dense representations rich in contextual information. To address the class imbalance, they used SMOTE (Synthetic Minority Oversampling Technique) to balance the training dataset.Embedded data warmed-up a CatBoost classifier for the reason that it is adept at dealing with categorical nearest neighbor features and unbalanced data sets. For evaluation, the team applied a 90:10 train-validation split and macro-averaged metrics were employed to allow for a comprehensive performance appraisal This method effectively combines the advantages of pre-trained embeddings and a powerful gradient boosting model, yielding accurate multi-class classification.

CUET_NetworkSociety team (Babu et al., 2025) employed a transformer-based approach using the 'bert-base-multilingual-cased' model for text classification. The data preprocessing includes normalization and label encoding. The team utilized the Hugging Face 'Trainer' class for fine-tuning with tokenized inputs, optimized hyperparameters, and mixed precision ('fp16') was implemented to enhance computational efficiency during training.

Walter White team utilized the Indic BERT model, which is well-suited for code-mixed data and effectively handles Tamil-specific linguistic features. During the preprocessing stage, the team replaced emojis with their corresponding textual descriptions but excluded those irrelevant to the context (e.g., the kite emoji). They also removed newline characters, hashtags, and normalized spaces for consistency. For tokenization, they opted for the Trivial Tokenizer, as it is compatible with both Indic BERT and the Tamil language.

YenCS team implemented a multi-step approach to text classification. Initially, the text data is preprocessed by cleaning and tokenizing it. Then, word embeddings are generated using a pre-trained word2vec model. Three different deep learning models are trained: a Convolutional Neural Network (CNN) with a GRU layer, an LSTM model, and an LSTM model with an added GRU layer. These models are then combined using a stacking ensemble technique, where the predictions of the individual models serve as input features for a meta-model (RandomForestClassifier). Finally, the meta-model makes the final prediction, aiming to improve the overall classification accuracy compared to using any single model alone. The process is further enhanced by using early stopping and hyperparameter tuning to optimize model performance.

ARINDASCI team performed political sentiment classification using a multi-step machine learning pipeline. Initially they preprocessed the data by removing the noise like special characters, URLs, and whitespaces. Then they tokenized and used pre-trained embeddings (e.g., fastText or TamilBERT) to capture the semantic informations For classification, the team experimented with various models, including traditional machine learning algorithms like Logistic Regression and advanced deep learning models such as LSTMs and Transformer-based architectures. The system achieved a macro-F1-score of 0.0727on the test set.

5 Results and Discussion

There was a total of 139 people who registered for this shared task, and 25 teams submitted their results. The ranking for Tamil was determined based on the macro F1-score, as shown in Table 2. The Synapse team secured first place with an F1-score of 0.377 by fine-tuning the IndicBERTv2-MLM-Back-TLM encoder-based LLM, leveraging IndicCorp v2 and Samanantar datasets. The KCLR team followed closely in second place, achieving a score of 0.371 with a transformer-based deep learning model enhanced through diverse embedding techniques. The byteSizedLLM team ranked third with an F1-score of 0.349, employing a hybrid approach that integrated a customized attention BiLSTM network with a fine-tuned XLM-RoBERTa base model.

Table 2: **Task: Tamil Rank list**

Team Name	F1-score	Rank
Synapse (KP et al., 2025)	0.3773	1
KCRL (Mia et al., 2025)	0.3710	2
byteSizedLLM	0.3497	3
Eureka-CIOL (Eram et al., 2025)	0.3187	4
Wictory (K et al., 2025)	0.3115	5
MNLP	0.3026	6
Nova Spark	0.3001	7
Team_Catalysts (Shanmugavadivel et al., 2025a)	0.2933	8
Lowes	0.2908	9
abhay43	0.2904	10
GS	0.2835	11
JAS	0.2796	12
SentiTamil	0.2769	13
CrewX	0.2759	14
AnalysisArchitects (Jayaraman et al., 2025)	0.2747	15
Beyond_tech (Shanmugavadivel et al., 2025b)	0.2736	16
CUET_Novice (Barua et al., 2025)	0.2728	17
KSK (M et al., 2025)	0.2654	18
QuanNguyen	0.2613	19
Team_Luminaries_0227	0.2530	20
VKG	0.2526	21
CUET_NetworkSociety (Babu et al., 2025)	0.2178	22
WalterWhite	0.1554	23
YenCS	0.1333	24
ARINDASCI_Tamil	0.0727	25

6 Conclusion

The "Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments" shared task provided valuable insights into the classification of Tamil political comments from social media. As part of the DravidianLangTech@NAACL workshop, this task challenged participants to categorize comments into seven predefined classes using diverse machine learning, deep learning, and natural language processing approaches. With 25 participating teams, model performance was assessed using the macro-F1 score. Given the small dataset size, few-shot and zero-shot learning strategies could enhance model efficiency. Furthermore, integrating Explainable AI (XAI) techniques can improve transparency and interpretability, fostering trust in model predictions and advancing sentiment analysis for low-resource languages like Tamil.

Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2), supported in part of Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

S Anbukkarasi and S Varadhaganapathy. 2020. Analyzing sentiment in Tamil tweets using deep neural network. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 449–453. IEEE.

D Anish and V Sumathy. Sentiment extraction for Tamil political.

Tofayel Ahmmed Babu, MD Musa Kalimullah Ratul, Sabik Aftahee, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET_NetworkSociety@DravidianLangTech 2025: A Transformer-Driven Approach to Political Sentiment Analysis of Tamil X (Twitter) Comments. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Arupa Barua, Md Osama, and Ashim Dey. 2025. CUET_Novice@DravidianLangTech 2025: A Bi-GRU Approach for Multiclass Political Sentiment Analysis of Tamil Twitter (X) Comments. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Irene V Blair. 2002. The malleability of automatic stereotypes and prejudice. *Personality and social psychology review*, 6(3):242–261.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596*.

K Devasena, M Sarika, and J Shana. 2022. Predicting Tamil nadu election 2021 results using sentimental analysis before counting. In *Proceedings of the International Conference on Computational Intelligence and Sustainable Technologies: ICoCIST 2021*, pages 279–289. Springer.

Patricia G Devine. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, 56(1):5.

Enjamamul Haque Eram, Anisha Ahmed, Sabrina Afroz Mitu, and Azmine Toushik Wasi. 2025. Eureka-CIOL@DravidianLangTech 2025: Using Customized BERTs for Sentiment Analysis of Tamil Political Comments. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *arXiv preprint arXiv:2305.08283*.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. *arXiv preprint arXiv:2106.11410*.

Alfred Hermida, Fred Fletcher, Darryl Korell, and Donna Logan. 2012. Share, like, recommend: Decoding the social media news consumer. *Journalism studies*, 13(5-6):815–824.

Abirami Jayaraman, Aruna Devi Shanmugam, Dharunika Sasikumar, and Bharathi B. 2025. AnalysisArchitects@DravidianLangTech 2025: Machine Learning Approach to Political Multiclass Sentiment Analysis of Tamil. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kenneth Joseph and Jonathan H Morgan. 2020. When do word embeddings accurately reflect surveys on our beliefs about people? *arXiv preprint arXiv:2004.12043*.

Nithish Ariyha K, Eshwanth Karti T R, Yeshwanth Balaji A P, Vikash J, and Sachin Kumar S. 2025. Wictory@DravidianLangTech 2025: Political Sentiment Analysis of Tamil X(Twitter) Comments using LaBSE and SVM. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Suriya KP, Durai Singh K, Vishal A S, Kishor S, and Sachin Kumar S. 2025. Synapse@DravidianLangTech 2025: Multiclass Political Sentiment Analysis in Tamil X (twitter) Comments: Leveraging Feature Fusion of IndicBERTv2 and Lexical Representations. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

S Kumar, V Balachandran, L Njoo, A Anastasopoulos, and Y Tsvetkov. 2022. Language generation models can cause harm: so what can we do about it. *An actionable survey. CoRR abs/2210.07700*.

Anna Sophie Kümpel, Veronika Karnowski, and Till Keyling. 2015. News sharing in social media: A review of current research on news sharing users, content, and networks. *Social media+ society*, 1(2):2056305115610141.

Thissyakkanna S M, Kalaivani K S, Sanjay R, and NIRENJHANRAM S K. 2025. KEC_AI_KSK@DravidianLangTech 2025: Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments Using Incremental Learning. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2020. JUNLP@ Dravidian-CodeMix-FIRE2020: Sentiment classification of code-mixed tweets using bi-directional RNN and language tags. *arXiv preprint arXiv:2010.10111*.

Md Ayon Mia, Fariha Haq, Md. Tanvir Ahammed Shawon, Golam Sarwar Md. Mursalin, and MUHAMMAD IBRAHIM KHAN. 2025. KCRL@DravidianLangTech 2025: Multi-View Feature Fusion with XLM-R for Tamil Political Sentiment Analysis. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Eni Mustafaraj and Panagiotis Takis Metaxas. 2011. What edited retweets reveal about online political discourse. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Lee Rainie, Aaron Smith, Kay Lehman Schlozman, Henry Brady, Sidney Verba, et al. 2012. Social media and political engagement. *Pew Internet & American Life Project*, 19(1):2–13.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of Tamil

code-mixed data. *Computer Speech & Language*, 76:101407.

Kogilavani Shanmugavadivel, Malliga Subramanian, Subhadevi K, Sowbharanika Janani Sivakumar, and Rahul K. 2025a. Team_Catalysts@DravidianLangTech-NAACL 2025: Leveraging Political Sentiment Analysis using Machine Learning Techniques for Classifying Tamil Tweets. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Sanjai R, Mohammed sameer, and Motheeswaran K. 2025b. Beyond_Tech@DravidianLangTech 2025: Political Multiclass Sentiment Analysis using Machine Learning and Neural Network. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

A Sharmista and Dr M Ramaswami. 2020. Sentiment analysis on Tamil reviews as products in social media using machine learning techniques: A novel study. *Madurai Kamaraj University Madurai-625*, 21.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on industrial and information systems (ICIIS)*, pages 320–325. IEEE.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 178–185.

Alcides Velasquez. 2012. Social media and online political discussion: The effect of cues and informational cascades on participation in online political communities. *New Media & Society*, 14(8):1286–1303.

Ekaterina Zhuravskaya, Maria Petrova, and Ruben Enikolopov. 2020. Political effects of the internet and social media. *Annual review of economics*, 12(1):415–438.

Overview of the Shared Task on Fake News Detection in Dravidian Languages-DravidianLangTech@NAACL 2025

Malliga Subramanian[1], Premjith B[2], Kogilavani Shanmugavadivel[1], Santhiya Pandiyan[1], Balasubramanian Palani[3], Bharathi Raja Chakravarthi[4]
[1]Kongu Engineering College, Tamil Nadu, India,
[2]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India,
[3]Indian Institute of Information Technology Kottayam, Kerala, India,
[4]School of Computer Science, University of Galway, Ireland

Abstract

Detecting and mitigating fake news on social media is critical for preventing misinformation, protecting democratic processes, preventing public distress, mitigating hate speech, reducing financial fraud, maintaining information reliability, etc. This paper summarizes the findings of the shared task "Fake News Detection in Dravidian Languages—DravidianLangTech@NAACL 2025." The goal of this task is to detect fake content in social media posts in Malayalam. It consists of two subtasks: the first focuses on binary classification (Fake or Original), while the second categorizes the fake news into five types—False, Half True, Mostly False, Partly False, and Mostly True. In Task 1, 22 teams submitted machine learning techniques like SVM, Naïve Bayes, and SGD, as well as BERT-based architectures. Among these, XLM-RoBERTa had the highest macro F1 score of 89.8%. For Task 2, 11 teams submitted models using LSTM, GRU, XLM-RoBERTa, and SVM. XLM-RoBERTa once again outperformed other models, attaining the highest macro F1 score of 68.2%.

1 Introduction

In the modern age, information is spreading rapidly across the world, and the quality and truthfulness of the news affect society. This fast, rapid connectivity democratizes access to knowledge and information. However, it has also created a possibility for the proliferation of fake news and misinformation. Therefore, detecting and mitigating the fake and misinformation has become critical. Detecting fake news is a complex task due to the structure of the sentence. Unlike tasks like hate speech detection and sentiment analysis, where we have overt words/phrases that explain the meaning of the sentence, fake news doesn't contain such words, whereas it mimics the legitimate content.

Detecting fake news from low-resource languages like Malayalam is even more challenging due to linguistic diversity and resource limitations (Raja et al., 2024, 2023b). In addition, the presence of code-mixed text (Coelho et al., 2023) and the requirement of fine-tuning pre-trained models (Raja et al., 2023a) pose other challenges.

The goal of the shared task DravidianLangTech@NAACL 2025 is to address the difficulties in Malayalam fake news detection. This paper presents an overview of the submissions to this shared task. This task has two subtasks. The first task is to identify whether a given news item is fake or not, and the second task is about categorizing news into different fake categories (Subramanian et al., 2024). As a part of this task, we curated our own dataset. We gathered news from various fact-checking websites in Malayalam.

2 Related Works

Fake news detection and categorization are important tasks in languages like Malayalam due to the rapid spread of misinformation. Various approaches, including machine learning, deep learning, and transformer-based models for feature extraction as well as classification.

Machine learning algorithms such as random forest, support vector machine (SVM), logistic regression, and naive Bayes were widely used for detecting fake news and categories of fake news (Bade et al., 2024; Osama et al., 2024; Devika et al., 2024). Deep learning models also find success in this task. Long Short-Term Memory (LSTM) (Zamir et al., 2024) and Bidirectional LSTM (BiLSTM) models achieved a macro F1 score of 0.78. Convolutional Neural Networks (CNN) (Osama et al., 2024) were also employed for this task. Recently, researchers used XLM-RoBERTa (Malliga et al., 2023; Raja et al., 2023a), MuRIL (Farsi et al., 2024), m-BERT (Osama et al., 2024), and

Malayalam BERT (Rahman et al., 2024; Tabassum et al., 2024). XLM-RoBERTa models achieved an F1 score of 0.87 and 0.90 in these tasks, whereas MuRIL-based models achieved an F1 score of 0.86 for the fake news detection task. Models built using m-BERT achieved similar performance with an F1 score of 0.85. Malayalam-BERT models achieved significant improvement in the categorization of fake news into different classes, with scores of 0.88 and 0.87.

3 Task description

3.1 Task 1

This task's objective is to determine if a particular social media text is original or fake; these data were sourced from numerous social media sites, including Facebook, Twitter, and others. The shared task's goal is to categorize a social media comment as either original or fake news. The classification of this task takes place at the comment/post level. The participant-submitted methods ought to classify a YouTube comment as either original or fake news.

3.2 Task 2

The primary objective of Task 2 is to classify fake news into different categories. In this task, we consider four classes of fake news, namely, false, mostly false, partly false, and half true. This classification helps people understand how much they have to rely on a specific news source to make their own decisions.

4 Dataset description

4.1 Task 1

The objective is to classify news items into 'Fake' and 'Original' categories. The dataset for this task comprises 1,599 training instances for 'Fake' and 1,658 for 'Original,' with respective testing sets of 507 and 512 instances nd development sets of 406 and 409 instances. Table 1 provides the number of data points in the training, development, and testing sets as well as the class-wise distribution

4.2 Task 2

In this task, the dataset was curated to contain different fake categories of Malayalam news, rather than classifying news into either fake or benign categories (Devika et al., 2024). We collected the news and their corresponding annotations from various fact-checking websites in Malayalam. We prepared a set of keywords to search for the news and identify their categories. To validate the authenticity of the annotations, we cross-checked them with multiple fact-checking tools. We provided train and test data for the participants of the shared task. Initially, we provided annotated training data for model building, and later, we provided test data without labels. Table 2 provides the number of data points in the training and testing sets as well as the class-wise distribution. The data is highly imbalanced, and the majority of the data in both the training and testing sets belong to the false category.

Data	Class	Count	Total
Train	Fake	1,599	3,257
	Original	1,658	
Development	Fake	406	815
	Original	409	
Test	Fake	507	1,019
	Original	512	

Table 1: Distribution of the data for Task 1

Data	Class	Count	Total
Train	False	1,386	1,900
	Mostly False	295	
	Partly False	57	
	Half True	162	
Test	False	100	200
	Mostly False	56	
	Partly False	7	
	Half True	37	

Table 2: A table explaining the distribution of the data in Train and Test datasets in Task 2

5 Methodology of participants

5.1 Task 1

Task 1 received 122 registrations. However, twenty-one teams actively participated and implemented their models. They tested the performance of the proposed models using the given fake news dataset, and the results are shown in Table 3.

5.1.1 Bytesizedllm

The team "bytesizedllm" (Manukonda and Kodali, 2025) developed an automatic fake news detection (FND) framework that uses a transformer-based fine-tuned XLM-RoBERTa model to lever-

age the strengths of both contextualized embeddings and sequential modeling. The transformer layer, integrated on top of the fine-tuned embeddings, further captures sequential dependencies, making the model highly effective for multilingual and transliteration-heavy tasks. The model has achieved the highest macro F1 score of 0.898 among all the other models proposed by other teams.

5.1.2 CUET_NLP_MP_MD

The team (Kabir et al., 2025) designed an FND system that combines multiple models, including Malayalam BERT, XLM-R, and Sarvamai/Sarvam-1 for contextual embedding and a majority voting classifier to detect fake news. This ensemble method leverages the strengths of each individual model to enhance performance and robustness. Hence, the model achieves a macro F1 score of 0.893 on the test data.

5.1.3 Awy

The team has employed a novel FND framework that consists of a mixture of multilingual models for contextual embedding and LLMs for emotion extraction to detect fake news effectively. The model has achieved a macro F1 score of 0.889.

5.1.4 Nayel

A machine learning-based system has been developed by the team 'Nayel' (Nayel et al., 2025), which integrates TF-IDF and n-grams as a feature extraction approach and sends the extracted features to ML-based classifiers such as SVM, SGD, Naive Bayes, and the Multi-Voting ensemble method to identify fake news. The highest macro F1 score of 0.875 is obtained by the ensemble model.

5.1.5 KCRL

The team (Haq et al., 2025) has implemented a text classification approach utilizing the XLM-RoBERTa base model augmented with a multi-pooling strategy. The methodology incorporates three distinct pooling mechanisms: CLS token extraction, mean pooling, and max pooling to capture comprehensive contextual representations from the input sequences. This unified pooling mechanism, enhanced by adaptive thresholding optimization, enables more robust classification by leveraging different semantic perspectives of the input text. Hence, their proposed model has achieved a macro F1 score of 0.874.

5.1.6 CUET-NLP_Big_O

The team (Sakib et al., 2025) has employed the XLM-RoBERTa (XLMR) large model, a multilingual transformer-based architecture, to classify social media text as either "Fake" or "Original". The model is tested on the dataset and achieves a macro F1 score of 0.874.

5.1.7 Celestia

The team (Noor et al., 2025) has designed an FND system that employs different embedding techniques and various ML and DL algorithms to detect fake news. The main advantage of this work includes the indic-transliteration library to create a consistent language format, English to Malayalam. The model achieves a macro F1 score of 0.859 on the test data.

5.1.8 MNLP

The team has developed an FND model that explores different deep learning-based models to identify fake news. The model has achieved a macro F1 score of 0.858.

5.1.9 CIC_NLP

The team (Achamaleh et al., 2025) has developed a novel FND framework that utilizes multilingual BERT (mBERT) for contextual word embedding for Tamil and Malayalam languages, and then the extracted features are sent to classifiers to detect fake news. The model achieves a macro F1 score of 0.853 on the unseen test dataset.

5.1.10 NLP_goats

The team (V K et al., 2025) implemented an automatic FND system that uses a multilingual BERT (mBERT) model for efficient fake news detection in Malayalam. These features make the model versatile and a very efficient solution for fake news detection in the Malayalam language. The model is tested on the dataset and achieves a macro F1 score of 0.839.

5.1.11 Necto

The team has utilized Sentence BERT, a fine-tuned model on the given data with a binary classification head for the classification downstream task. The sentence-level embeddings of the given text and the size of the model are small so that the training time of the model is faster in any system. Hence, the model achieves a macro F1 score of 0.832 on the test data.

5.1.12 Lowes

The team has developed an FND system that utilizes mBERT and various LLM-based approaches to detect fake news. The model has achieved a macro F1 score of 0.826.

5.1.13 Lemlem

The team 'Lemlem' has employed the pre-trained multilingual transformer model named mBERT for the word embedding that captures contextual relationships between words to detect fake news. This model is particularly suitable for multilingual text processing as it can handle diverse scripts and linguistic features effectively. A classification head was added to the BERT base model, which outputs probabilities for the predefined classes of fake news. Hence, the model achieves a macro F1 score of 0.823 on the test data.

5.1.14 Data_drifters

For this task, the team (Shanmugavadivel et al., 2025a) has employed a comprehensive methodology combining traditional machine learning models, embedding techniques, and advanced transfer learning models to achieve robust text classification. The team utilized four base models: Random Forest, Support Vector Machine (SVM), Logistic Regression, and Multinomial Naive Bayes, leveraging their diverse strengths in classification tasks. Two classical count-based techniques, such as TF-IDF and Count Vectorizer, are used to convert words into vectors. In addition, mBERT and XLNet are utilized for their exceptional ability to understand contextual semantics and multilingual text. The model is tested on the dataset and achieves a macro F1 score of 0.814.

5.1.15 ST_1 CIOL

The team (Anik et al., 2025) has designed a novel FND model that employs a multilingual encoder to effectively encode the text into meaningful embeddings. These embeddings were then fed into a Multi-Layer Perceptron (MLP) model for training, enabling the prediction of sentiment classes. The team has adopted a balance-aware modeling approach, actively tracking the best-performing model throughout the training process to ensure optimal performance. Finally, the best model is utilized for generating predictions on the test set and achieves a macro F1 score of 0.814.

5.1.16 Fact_fusion

The team has implemented a multilingual pipeline for detecting fake news for the Dravidian languages based on a systematic methodology. TF-IDF (Term Frequency-Inverse Document Frequency) was used for extracting features from the textual data, considering the term's significance but minimizing noise. The logistic Regression model has achieved a better macro F1 score of 0.803 when compared to other models.

5.1.17 YenCs

The team (Gowda and Hegde, 2025) has employed a novel automatic FND system that uses four different deep learning models (BiRNN, DNN, GRU, LSTM + RNN) with pre-trained word embeddings and combines their predictions using a weighted average based on validation accuracies. These models and the ensemble model were evaluated using metrics like accuracy and F1-score and achieved a macro F1-score of 0.792 on the unseen test data.

5.1.18 Blue_ray

The team (Shanmugavadivel et al., 2025b) has developed an FND system to classify Malayalam news into two categories: Original and Fake. After text pre-processing, features are extracted using TF-IDF to capture significant patterns in the text. Various machine learning models, such as Logistic Regression, Random Forest, and SVM, are trained on these features to predict the labels. The model achieves a macro F1-score of 0.790 on the unseen test dataset.

5.1.19 CIC

The team has developed a novel model to detect fake news. First, the proposed model performed tokenization and other pre-processing with indic_nlp method and then applied feature extraction using a fine-tuned mBERT model for training and prediction. The model is tested on the dataset and achieved the macro F1 score of 0.659.

5.1.20 DLRG

For the fake news classification in Malayalam, the team implemented an FND system that employs TF-IDF to convert text data into numerical features, highlighting important words. Then, Passive passive-aggressive classifier (PAC) is used to classify the TF-IDF transformed data into fake or real news. In addition, a Voting Classifier is utilized to combine predictions from multiple classifiers to

enhance accuracy. Hence, the model achieves a macro F1-score of 0.473 on the test data.

5.1.21 CUET_ChiSquare

The team has designed a novel automatic FND system that utilizes a transformer-based approach leveraging XLM-RoBERTa, a multilingual pre-trained transformer model, fine-tuned for binary classification. The system's capability to generalize across varied linguistic structures and its efficient handling of imbalanced data make it particularly noteworthy for tasks involving low-resource and diverse language datasets. The model is achieved a macro F1-score of 0.334.

5.2 Task 2

Similar to Task 1, 122 teams registered for Task 2. However, only 11 teams submitted the predictions for the test data shared with the participants. The rank list for this task is shown in table 4. The following are the descriptions of the systems submitted by the participants.

5.2.1 byteSizedLLM

This team (Kodali and Manukonda, 2025) employed an advanced hybrid methodology, combining a customized BiLSTM network with a fine-tuned XLM-RoBERT base model to leverage the strengths of both contextualized embeddings and sequential modeling. The XLM-RoBERTa base model was fine-tuned using masked language modeling (MLM) on a carefully curated subset of the AI4Bharath dataset designed to enhance its multilingual contextual understanding. The dataset had original data, fully transliterated text, and partially transliterated data, with 20% to 70% of words randomly transliterated. This was done to add transliteration-based diversity. This method lets the model learn strong cross-lingual representations and adjust to different transliteration patterns that are common in collections of texts written in more than one language. The BiLSTM layer, which is added on top of the fine-tuned embeddings, captures even more sequential dependencies. This makes the model very good at tasks that require a lot of transliteration and more than one language.

5.2.2 YenCS

In this submission, the team (Gowda and Hegde, 2025) pre-processed the input text. The preprocessing step includes cleaning and tokenisation using Keras's Tokeniser. A pre-trained fastText model transformed the cleaned words into embeddings. The team trained three different models: a Convolutional Neural Network (CNN) with a GRU layer, an LSTM model, and an LSTM-GRU hybrid model. They subsequently use the predictions from these models as features in a stacking ensemble. A random forest classifier serves as the meta-learner, trained on the stacked predictions to produce the final classification. The team evaluated the effectiveness of both individual models and groups of models through accuracy and classification reports.

5.2.3 Fact Fusion

This team used a machine learning pipeline for the fake news classification task. This pipeline begins with text preprocessing, which removes noise like punctuation and extra white spaces. They used the term frequency-inverse document frequency (TF-IDF) for transforming the input text into embeddings. Unigrams and bigrams were considered for defining the features, and they restricted the vocabulary size to 5000 words. They performed the classification using a logistic regression classifier. They optimised the model training by fine-tuning the hyperparameter "max_iter."

5.2.4 Lowes

This team finetuned Malayalam BERT and also tried other LLM-based approaches. They also tried LLM-based synthetic data generation for this task.

5.2.5 KCRL

The team (Haq et al., 2025) developed a text classification system using the XLM-RoBERTa transformer model and improved it with a full pooling strategy and a data-balancing method. This method uses three different pooling methods—CLS token, mean pooling, and max pooling—to capture different parts of textual representations. To address class imbalance issues, they implemented an oversampling approach for minority classes, targeting a balanced distribution across all classes. Before the final classification, the concatenated pooled features go through a dense layer transformation. This lets the model use both global and local semantic features while keeping training levels even across all classes.

5.2.6 NLP_goats

The team (V K et al., 2025) developed a model that begins with preprocessing of the Malayalam text. They then encode the dataset labels and address imbalances through oversampling techniques.

The team trains a multilingual BERT model for multiclass classification.

5.2.7 MNLP

This team used deep learning-based models for classification.

5.2.8 Akatsuki-CIOL

This team (Anik et al., 2025) used a variety of encoders, such as Indic-specific and language-specific models, to get meaningful text embeddings that fit the multilingual nature of the data. They then processed these embeddings using the multilayer perceptron (MLP). It was the classification layer that predicted the corresponding classes. To ensure robust performance, the team adopted a balance-aware modeling approach, actively tracked and selected the best-performing model throughout the training process. Then, we used the chosen model to make predictions on the test set. This gave us a complete and flexible way to solve the multilingual sentiment classification problem.

5.2.9 Data_Drifters

For this task, the team (Shanmugavadivel et al., 2025a) employed a comprehensive methodology combining traditional machine learning models, embedding techniques, and advanced transfer learning models to achieve robust text classification. They used four base models: random forest, support vector machine (SVM), logistic regression, and multinomial logistic regression, leveraging their diverse strengths in classification tasks. To process text data, they implemented two embedding techniques: TF-IDF and Count Vectorizer, ensuring effective feature extraction and representation. Additionally, they incorporated two state-of-the-art transfer learning models, mBERT and XLNet.

5.2.10 Blue_Ray

The team implemented a multi-class classification system to categorize fake Malayalam news. The methodology involved preprocessing the text data by cleaning it and removing stopwords to ensure better feature representation. Feature extraction transformed the text data into the numerical format. They then split the processed data into training and testing subsets. Various machine learning and deep learning models were employed to train the data, and their performance was evaluated.

5.2.11 Cognitext

The team (Alladi and B, 2025) used a deep learning model to classify fake news articles. They first cleaned the text data by removing URLs, special characters, punctuation, and numbers, and then converted it to lowercase. They tokenized the cleaned text using Keras' Tokenizer and padded the sequences to ensure a uniform input length. The model architecture is made up of an embedding layer that stores word representations, an LSTM layer that tracks how events depend on each other, and a dense layer that uses softmax activation to sort words into multiple groups. They trained the model for five epochs using categorical cross-entropy loss and the Adam optimizer.

Team	Macro F1-score	Rank
bytesizedllm (Manukonda and Kodali, 2025)	0.898	1
CUET_NLP_MP_MD (Kabir et al., 2025)	0.893	2
One_by_zero (Chakraborty et al., 2025)	0.892	3
Awy	0.889	4
Nayel (Nayel et al., 2025)	0.875	5
KCRL (Haq et al., 2025)	0.874	6
CUET-NLP_Big_O (Sakib et al., 2025)	0.874	6
Celestia (Noor et al., 2025)	0.859	7
MNLP	0.858	8
CIC_NLP (Achamaleh et al., 2025)	0.853	9
NLP_goats (V K et al., 2025)	0.839	10
Necto	0.832	11
Lowes	0.826	12
Lemlem	0.823	13
Data_drifters (Shanmugavadivel et al., 2025a)	0.814	14
ST_1 CIOL	0.814	14
Fact_fusion	0.803	15
YenCs (Gowda and Hegde, 2025)	0.792	16
Blue_ray (Shanmugavadivel et al., 2025b)	0.790	17
CIC	0.659	18
DLRG	0.473	19
Technovators	0.387	20
CUET_ChiSquare	0.334	21

Table 3: Rank list of Task 1: Detecting fake news in Malayalam

Team	Macro F1-score	Rank
KCRL (Haq et al., 2025)	0.6283	1
byteSizedLLM (Kodali and Manukonda, 2025)	0.5775	2
NLP_goats (V K et al., 2025)	0.5417	4
Data_Drifters(Shanmugavadivel et al., 2025a)	0.5029	5
lowes	0.2902	6
YenCS (Gowda and Hegde, 2025)	0.2696	7
Blue_Ray (Shanmugavadivel et al., 2025b)	0.2631	8
Akatsuki-CIOL (Anik et al., 2025)	0.1978	11
Cognitext (Alladi and B, 2025)	0.1667	14
Fact-Fusion	0.1667	14
MNLP	0.1667	14

Table 4: Rank list of Task 2: Classification of fake news into various categories

This task saw a significant difference in the model performance. In Task 1, top teams achieved macro F1 scores greater than 0.89, whereas the lower-ranked teams attained scores around 0.33. The trend is similar in Task 2, too. The first-ranked

team scored an F1 score of 0.6283, and the bottom-ranked teams scored only 0.1667. Most of these differences in performance can be traced back to the approaches that the teams devised with class imbalance and data augmentation, feature extraction, and model architecture.

Class imbalance is a primary challenge pertaining to this task, especially in Task 2. Some of the top-performing teams effectively addressed this issue by employing oversampling algorithms at the feature level. In addition, teams using the adaptive thresholding optimization approach ensured that the models did not overfit to the majority classes. The lower-ranked teams did not employ any mechanism to address the class imbalance issue and hence resulted in the poor generalization of the minority class data samples.

The selection of the model architectures played a pivotal role in the performance. The majority of the top-ranked teams leveraged the transformer-based architectures, which provided better context understanding of the data compared to the traditional machine learning classifiers and feature extraction approaches. The transformer-based models excelled because of their multilingual capabilities, making them effective for data in low-resource languages like Malayalam. In addition, these models provide better learning and contextual token embeddings due to their multilingual capability. Traditional feature extraction models such as TF-IDF, bag-of-words, and n-gram-based representations struggled to capture the deep contextual and semantic relationships in fake news content, leading to suboptimal performance. In addition, data augmentation and the use of ensemble models significantly improve their performance. The availability of the computational resources played a major role in determining the performance of the models. The teams who fine-tuned the transformer models using the task achieved better scores compared to the models that did not use it.

For the most part, the best teams use data enhancement techniques, transformer-based architectures, ensemble methods, and computing resources well. The most effective models incorporated oversampling for class balance, transliteration-aware augmentation, hybrid architectures combining transformers and LSTMs, and multi-layered pooling strategies. In contrast, teams that relied on simpler machine learning models, failed to address class imbalance, or lacked data augmentation strategies struggled to achieve competitive results. These findings highlight the importance of adaptive learning techniques and advanced model enhancement strategies for tackling complex NLP tasks like fake news detection.

6 Conclusion

This paper presents a summary of the shared task "Fake News Detection in Dravidian Languages - DravidianLangTech@NAACL 2025," which focuses on the Malayalam language. The task provided an opportunity to assess the efficacy of several machine learning and deep learning algorithms in detecting fake news on social media. Transformer-based architectures, particularly XLM-RoBERTa, consistently outperformed traditional machine learning algorithms, with the highest macro F1-scores in both binary and multi-class classification tasks. These findings give the promise of advanced NLP models in handling fake news and emphasize the significance of continued research and model improvement to enhance accuracy.

Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2), supported in part of Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

Tewodros Achamaleh, Nida Hafeez, Mikiyas Mebraihtu, Fatima Uroosa, and Grigori Sidorov. 2025. CIC-NLP@DravidianLangTech 2025: Fake News Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shriya Alladi and Bharathi B. 2025. Cognitext@DravidianLangTech2025: Fake News Classification in Malayalam Using mBERT and LSTM. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Mahfuz Ahmed Anik, Md. Iqramul Hoque, Wahid Faisal, Azmine ToushikWasi, and Md Manjurul Ahsan. 2025. Akatsuki-CIOL@DravidianLangTech 2025: Ensemble-Based Approach Using Pre-Trained Models for Fake News Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on*

Speech, Vision, and Language Technologies for Dravidian Languages. Association for Computational Linguistics.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024. Social Media Fake News Classification Using Machine Learning Algorithm. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 24–29.

Dola Chakraborty, Shamima Afroz, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. One_by_zero@DravidianLangTech 2025: Fake News Detection in Malayalam Language Leveraging Transformer-based Approach. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sharal Coelho, Asha Hegde, G Kavya, and Hosahalli Lakshmaiah Shashirekha. 2023. Mucs@ dravidianlangtech2023: Malayalam fake news detection using machine learning approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 288–292.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From Dataset to Detection: A Comprehensive Approach to Combating Malayalam Fake News. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Salman Farsi, Asrarul Eusha, Ariful Islam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@ DravidianLangTech EACL2024: Fake News Detection in Malayalam Language Leveraging Fine-tuned MuRIL BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 173–179.

Anusha M D Gowda and Parameshwar R Hegde. 2025. YenCS@DravidianLangTech 2025: Integrating Hybrid Architectures for Fake News Detection in Low-Resource Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Fariha Haq, Md. Tanvir Ahammed Shawon, Md Ayon Mia, Golam Sarwar Md. Mursalin, and Muhammad Ibrahim Khan. 2025. KCRL@DravidianLangTech 2025: Multi-Pooling Feature Fusion with XLM-RoBERTa for Malayalam Fake News Detection and Classification. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Md Minhazul Kabir, Md. Mohiuddin, Kawsar Ahmed, and Moshiul Mohammed Hoque. 2025. CUET-NLP_MP@DravidianLangTech 2025: A Transformer and LLM-Based Ensemble Approach for Fake News Detection in Dravidian. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Rohith Gowtham Kodali and Durga Prasad Manukonda. 2025. byteSizedLLM@DravidianLangTech 2025: Fake News Detection in Dravidian Languages Using Transliteration-Aware XLM-RoBERTa and Attention-BiLSTM. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

S Malliga, Bharathi Raja Chakravarthi, SV Kogilavani, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, and Muskaan Singh. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 59–63.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@DravidianLangTech 2025: Fake News Detection in Dravidian Languages Using Transliteration-Aware XLM-RoBERTa and Transformer Encoder-Decoder. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Hamada Nayel, Mohammed Aldawsari, and Hosahalli Lakshmaiah Shashirekha. 2025. NAYEL@DravidianLangTech-2025: Character N-gram and Machine Learning Coordination for Fake News Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Syeda Alisha Noor, Sadia Anjum, Syed Ahmad Reza, and Md Rashadur Rahman. 2025. Celestia@DravidianLangTech 2025: Malayalam-BERT and m-BERT based transformer models for Fake News Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_GoodFellows@ DravidianLangTech EACL2024: A Transformer-Based Approach for Detecting Fake News in Dravidian Languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192.

Tanzim Rahman, Abu Raihan, Md Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@ Dra-

vidianLangTech EACL2024: Fake News Classification Using Malayalam-BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 223–228.

Eduri Raja, Badal Soni, and Sami Kumar Borgohain. 2023a. nlpt malayalm@ DravidianLangTech: Fake news detection in Malayalam using optimized XLM-RoBERTa model. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 186–191.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023b. Fake news detection in Dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2024. Fake news detection in Dravidian languages using multiscale residual CNN_BiLSTM hybrid model. *Expert Systems with Applications*, 250:123967.

Nazmus Sakib, Md. Refaj Hossan, Alamgir Hossain, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET-NLP_Big_O@DravidianLangTech 2025: A BERT-based Approach to Detect Fake News from Malayalam Social Media Texts. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Vishali K S, Priyanka B, and Naveen Kumar K. 2025a. KEC_AI_DATA_DRIFTERS@DravidianLangTech 2025: Fake News Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Aiswarya M, Aruna T, and Jeevaananth S. 2025b. BlueRay@DravidianLangTech-2025: Fake News Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the Second Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@ EACL 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_punctuators@ dravidianlangtech-eacl2024: Transformer-based approach for detection and classification of fake news in malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186.

Srihari V K, Vijay Karthick Vaidyanathan, and Thenmozhi Durairaj. 2025. NLP_goats@DravidianLangTech 2025: Detecting Fake News in Dravidian Languages: A Text Classification Approach. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

M Zamir, M Tash, Z Ahani, A Gelbukh, and G Sidorov. 2024. Tayyab@ dravidianlangtech 2024: detecting fake news in malayalam lstm approach and challenges. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 113–118.

Incepto@DravidianLangTech 2025: Detecting Abusive Tamil and Malayalam Text Targeting Women on YouTube

Luxshan Thavarasa
Dept. of Computer Sci. and Eng
University of Moratuwa
Colombo, Sri Lanka
luxshan.20@cse.mrt.ac.lk

Sivasuthan Sukumar
Dept. of Electrical Eng
University of Moratuwa
Colombo, Sri Lanka
sivasuthansukumar@gmail.com

Jubeerathan Thevakumar
Dept. of Computer Sci. and Eng
University of Moratuwa
Colombo, Sri Lanka
jubeerathan.20@cse.mrt.ac.lk

Abstract

This study introduces a novel multilingual model designed to effectively address the challenges of detecting abusive content in low-resource, code-mixed languages, where limited data availability and the interplay of mixed languages, leading to complex linguistic phenomena, create significant hurdles in developing robust machine learning models. By leveraging transfer learning techniques and employing multi-head attention mechanisms, our model demonstrates impressive performance in detecting abusive content in both Tamil and Malayalam datasets. On the Tamil dataset, our team achieved a macro F1 score of 0.7864, while for the Malayalam dataset, a macro F1 score of 0.7058 was attained. These results highlight the effectiveness of our multilingual approach, delivering strong performance in Tamil and competitive results in Malayalam.

1 Introduction

Social media platforms play an essential role in modern communication, information sharing, and entertainment. However, they have also become spaces where harmful behavior proliferates, particularly in the form of abusive language targeting women. This abuse, often rooted in societal biases and gender inequalities, can have severe psychological, social, and professional consequences for victims (Jane, 2020). Tackling this issue is critical to creating safer and more inclusive digital spaces.

This research focuses on detecting abusive content in comments, with particular emphasis on Tamil and Malayalam—two low-resource languages spoken in South India. Online abuse in these languages is a pressing concern, but the limited availability of linguistic resources and tools presents significant challenges for effective content moderation. To address this, we leverage existing datasets introduced by Priyadharshini et al. (2023, 2022), which include YouTube comments collected around controversial and sensitive topics where gender-based abuse is prevalent. These datasets are annotated with binary labels: Abusive and Non-Abusive.

We adopt a transfer learning approach by utilizing the outputs from the last hidden layer of XLM-RoBERTa (Conneau et al., 2019), incorporating multi-head attention mechanisms to improve classification performance. This approach is well-suited for handling text in Tamil and Malayalam, addressing the challenges associated with detecting abusive content in these low-resource languages. Our model can be accessed via PyPI[1], and the complete work is available on GitHub[2].

The remainder of this paper is organized as follows: we discuss related work in abusive language detection for low-resource languages, describe the datasets and methodology, and present the results and evaluation metrics. This work aims to contribute to research in abusive language detection while highlighting the challenges and opportunities in working with Tamil and Malayalam.

2 Related Work

Detecting abusive and offensive content in low-resource languages, such as Tamil and Malayalam, is a critical research area due to rising online hate speech.

Arora (2020) introduced a model for Tamil-English code-mixed hate speech detection, uti-

[1] https://pypi.org/project/dravida-kavacham/
[2] https://github.com/Luxshan2000/dravida-kavacham

lizing a pre-trained ULM-FiT to handle code-mixed complexities. Ziehe et al. (2021) fine-tuned XLM-RoBERTa for Hope Speech detection in English, Malayalam, and Tamil, highlighting transformer adaptability in resource-constrained settings. Language-specific models like MuRIL (Khanuja et al., 2021), IndicBERT (Kakwani et al., 2020), and multilingual XLM-RoBERTa have accelerated research in Tamil.

Priyadharshini et al. (2022) explored abusive comment detection in Tamil and Tamil-English datasets, evaluating Logistic Regression (LR), Linear SVM, RNNs, Vanilla LSTMs, and transformer models like mBERT, MuRIL BERT, and XLM-RoBERTa. MuRIL BERT excelled due to its specialized training.

Chakravarthi et al. (2023) examined fine-grained abusive comment detection on Tamil-English YouTube data using BiLSTM with Attention and transformers like MuRIL-LARGE and XLM-R, where MuRIL-LARGE performed well. Sreelakshmi et al. (2024) addressed hate speech detection in Kannada-English, Malayalam-English, and Tamil-English datasets, testing BERT, DistilBERT, LaBSE, MuRIL, and IndicBERT. MuRIL embeddings with an SVM (RBF kernel) performed consistently well. A cost-sensitive learning approach addressed data challenges.

Malliga Subramanian (2023) advanced abusive Tamil comment detection by fine-tuning adapter-based transformers on datasets from (Priyadharshini et al., 2022). Researchers tested mBERT, MuRIL, and XLM-RoBERTa, achieving F1 scores below 0.73 but demonstrating adapter-based techniques' promise.

Other efforts, such as Patankar et al. (2022), combined classical machine learning and deep learning. Transformers like MuRIL, XLM-RoBERTa, and mBERT performed better, reaffirming their suitability for code-mixed scenarios.

These studies emphasize transformer-based architectures, such as MuRIL and XLM-RoBERTa, in tackling abusive content detection in Indian languages. By improving language-specific modeling and exploring multimodal approaches, these efforts have laid a foundation for further advancements.

3 Dataset

We use two datasets for this research: the Tamil and Malayalam datasets from Priyadharshini et al. (2023, 2022). The Tamil dataset consists of 2790 (Non-Abusive ≈1425, Abusive ≈1375) samples in the training set, 598 samples in the development set, and 598 samples in the test set. The Malayalam dataset contains 2933 (Non-Abusive ≈1525, Abusive ≈1400)samples in the training set, 629 samples in the development set, and 629 samples in the test set. Both datasets are annotated with binary labels: Abusive and Non-Abusive.

Figures 1 and 2 illustrate the length distribution of the training datasets, while Figures 3 and 4 present the word clouds for the respective datasets.

4 Methodology

4.1 Data Processing

For data preprocessing, we performed several cleaning steps to ensure the quality and consistency of the dataset. First, URLs were removed from the comments to eliminate any potential noise. Special characters were also removed to standardize the text. Additionally, emojis were excluded from the dataset. Since emojis rarely indicate abuse or non-abuse, we removed them to ensure consistency.(Kovács et al., 2021)

4.2 Model Architecture

Our model is a multilingual design tailored to classify text in both Tamil and Malayalam. After preprocessing, the inputs are tokenized using the XLM-RoBERTa-base tokenizer, and sentence embeddings are generated using the XLM-RoBERTa-base model, which is adept at handling multilingual tasks.

To capture task-specific features effectively, the embeddings are passed through four multi-head attention layers. These layers allow the model to focus on critical aspects of the input relevant to the classification task. The outputs then flow through dense layers with LayerNorm, ensuring stability and efficient learning. Finally, a softmax layer generates the classification probabilities.

Figure 5 illustrates the detailed architecture of the model.

5 Experiments and Results

For both Tamil and Malayalam datasets, we performed a 70% training, 15% validation, and 15% test split. The model was trained for 30 epochs with a dropout rate of 0.3 and ReLU activation. Training the model took approximately 1 hour on a Tesla P100 GPU, which efficiently handled the task with its 16GB memory. We used the macro-average F1

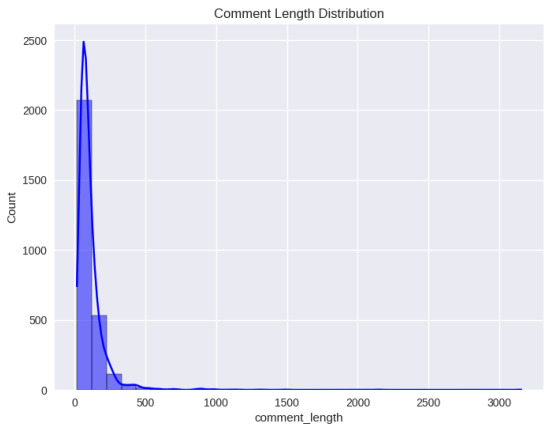


Figure 1: Distribution of comment lengths in Tamil language dataset.

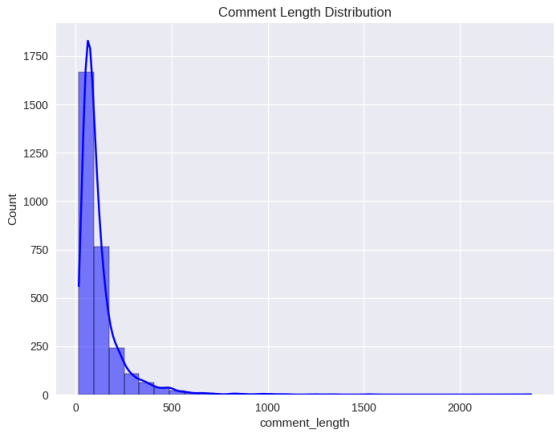


Figure 2: Distribution of comment lengths in Malayalam language dataset.

Figure 3: Word cloud representing all comments in Tamil.

score as the primary evaluation metric, ensuring that the model's performance was balanced across both classes (Abusive and Non-Abusive).

In the Tamil dataset, our team, Incepto, achieved an impressive macro F1 score of 0.7864, securing 3rd place out of 27 teams on the Dravidian-LangTech 2025 leaderboard(Rajiakodi et al., 2025). The performance was just 0.0019 behind the 1st rank team, CUET_Agile, which scored 0.7883. This demonstrates that our model was highly competitive and nearly matched the best performance on the leaderboard.

For the Malayalam dataset, Incepto ranked 4th out of 35 teams with a macro F1 score of 0.7058. The top team, Habiba A, G Agila, achieved a higher score of 0.7571, highlighting the competitive nature of the challenge.

These results underscore the effectiveness of our multilingual model, which demonstrated strong performance in detecting abusive content in Tamil and competitive results in Malayalam. While the model excelled in Tamil, there is room for optimization, especially for Malayalam, and future work can focus on improving performance further.

6 Conclusion

In this research, we tackled the task of detecting abusive language in Tamil and Malayalam by lever-

Figure 4: Word cloud representing all comments in Malayalam.

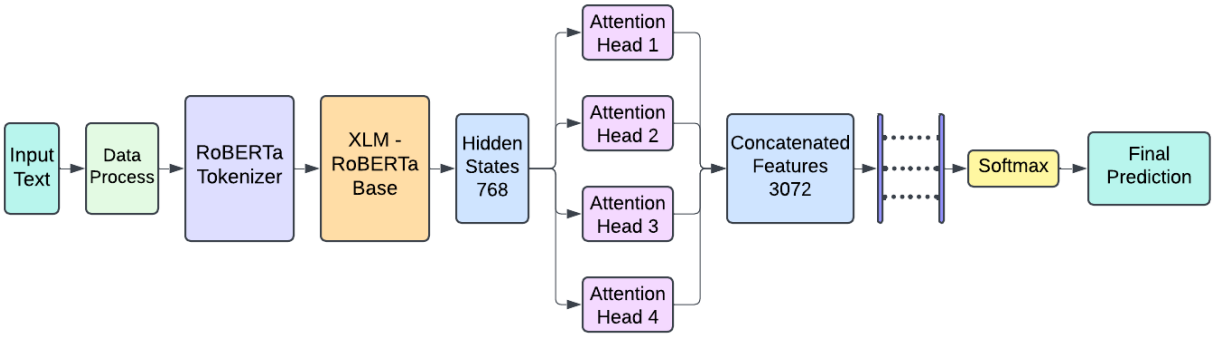


Figure 5: Proposed Model Architecture

aging a multilingual model, XLM-RoBERTa, augmented with multi-head attention mechanisms. Our approach delivered competitive results in Tamil (3rd rank) and highlighted challenges in Malayalam (4th rank), emphasizing the need for continued efforts in refining models for low-resource languages. To promote reproducibility and encourage further research, we published the model as a Python package on PyPI and made the source code publicly available. This contribution enables other researchers to replicate, analyze, and improve upon our work, fostering collaboration toward building safer and more inclusive online spaces for underrepresented language communities.

7 Limitations

One of the key limitations of this study is the inadequacy of annotated datasets for both Malayalam and Tamil, which affects model effectiveness and generalizability. The preprocessing and classification of these texts are particularly challenging due to the informal nature of social media language, which includes regional variations, shortened expressions, and a lack of grammatical structure. Additionally, while Malayalam and Tamil belong to the Dravidian language family, they differ significantly in morphology, syntax, and semantics, further complicating text analysis. Another major challenge is the absence of a proper tokenizer for both Tamil and Malayalam, making text processing and model training even more complex.

References

Gaurav Arora. 2020. Gauravarora@hasoc-dravidian-codemix-fire2020: Pre-training ulmfit on synthetically generated code-mixed data for hate speech detection. *Preprint*, arXiv:2010.02094.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Emma A Jane. 2020. Online abuse and harassment. *The international encyclopedia of gender, media, and communication*, 116.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

György Kovács, Pedro Alonso, and Rajkumar Saini. 2021. Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, 2(2):95.

Nandhini Subbarayan et al. On finetuning Adapter-based Transformer models for classifying Abusive Social Media Tamil Comments 22 February 2023 PREPRINT (Version 1) available at Research Square. Malliga Subramanian, Kogilavani Shanmugavadivel. 2023. On finetuning adapter-based transformer models for classifying abusive social media tamil comments. *SN Computer Science*, 1.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. Optimize$_prime$@$dravidianlangtech - acl2022 : Abusivecommentdetectionintamil$. *Preprint*, arXiv:2204.09675.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages (DravidianLangTech 2023)*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Stefan Ziehe, Franziska Pannach, and Aravind Krishnan. 2021. GCDH@LT-EDI-EACL2021: XLM-RoBERTa for hope speech detection in English, Malayalam, and Tamil. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 132–135, Kyiv. Association for Computational Linguistics.

Eureka-CIOL@DravidianLangTech 2025: Using Customized BERTs for Sentiment Analysis of Tamil Political Comments

Enjamamul Haque Eram, Anisha Ahmed, Sabrina Afroz Mitu, Azmine Toushik Wasi[†]
Shahjalal University of Science and Technology, Sylhet, Bangladesh
[†]**Correspondence:** azmine32@student.sust.edu

Abstract

Sentiment analysis on social media platforms plays a crucial role in understanding public opinion and the decision-making process on political matters. As a significant number of individuals express their views on social media, analyzing these opinions is essential for monitoring political trends and assessing voter sentiment. However, sentiment analysis for low-resource languages, such as Tamil, presents considerable challenges due to the limited availability of annotated datasets and linguistic complexities. To address this gap, we utilize a novel dataset encompassing seven sentiment classes, offering a unique opportunity to explore sentiment variations in Tamil political discourse. In this study, we evaluate multiple pre-trained models from the Hugging Face library and experiment with various hyperparameter configurations to optimize model performance. Our findings aim to contribute to the development of more effective sentiment analysis tools tailored for low-resource languages, ultimately empowering Tamil-speaking communities by providing deeper insights into their political sentiments. Our full experimental codebase is publicly available at: ciol-researchlab/NAACL25-Eureka-Sentiment-Analysis-Tamil

1 Introduction

Sentiment analysis is a crucial aspect of Natural Language Processing (NLP) that facilitates the categorization of textual opinions into various sentiment classes, such as positive, negative, neutral, and more. It has significant applications in understanding political discourse, enabling tasks such as forecasting election outcomes, analyzing public sentiment, and formulating targeted policies. Social media platforms like X (formerly Twitter) have gained immense popularity and have become a major hub for political discussions. In India, a substantial number of individuals actively use X to express their thoughts and opinions on political

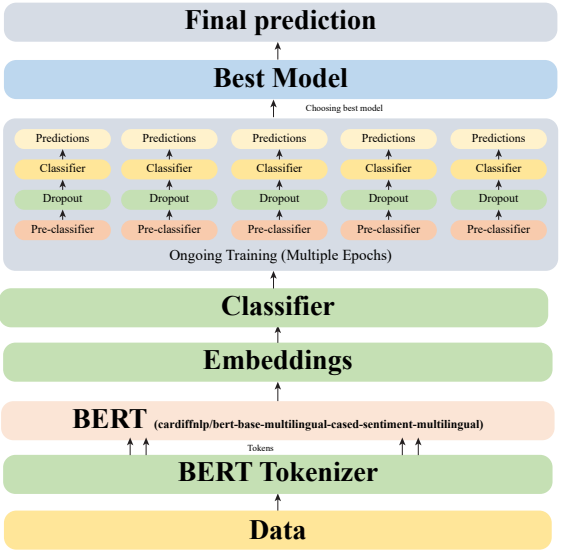


Figure 1: Model architecture, containing tokenizer, pre-trained model, classifier and other components

issues. As of 2018, the platform reported over 321 million monthly active users (Wang et al., 2012), with approximately 34.4 million users from India. Although this constitutes a small fraction of India's total population, the platform is extensively used by politicians, influencers, celebrities, and well-educated individuals whose tweets can significantly influence public sentiment. Their opinions can shape political narratives and sway the perspectives of their followers, highlighting the importance of structured sentiment analysis in this domain. Categorizing political content through sentiment analysis enables individuals to identify and engage with similar or opposing viewpoints, fostering a more organized exchange of information and enhancing public participation and awareness of political issues (Ansari et al., 2020).

Twitter data have been leveraged for various research applications, including sentiment analysis (Kouloumpis et al., 2011), stock market prediction (Bollen et al., 2011), trend detection (Mathioudakis and Koudas, 2010), information credibility assessment (Castillo et al., 2011), and event detection

(Becker et al., 2021). However, sentiment analysis in Tamil presents unique challenges due to the language's complex linguistic structure and diverse dialectical variations. While previous studies have explored sentiment analysis in Tamil, its application within the political discourse domain remains relatively underexplored. Addressing this gap, our study aims to develop effective sentiment analysis models tailored to Tamil political content, contributing to a deeper understanding of public opinion within this low-resource language setting.

Tamil-speaking individuals play a crucial role in shaping Indian politics, influencing public opinion and driving political discourse. Tamil, a prominent member of the Dravidian language family, is one of the oldest and most widely spoken languages in India, with a rich cultural and linguistic heritage. Despite its significance, there is a notable lack of dedicated resources and tools for analyzing political sentiment in Tamil. The Dravidian languages, including Telugu, Kannada, and Malayalam, share common linguistic features but also exhibit distinct characteristics, making them a unique challenge for NLP tasks (Chakravarthi et al., 2021, 2022, 2023, 2024). With the increasing prominence of social media as a platform for political discussions, this gap presents a significant challenge. Political opinions expressed on social media are inherently complex, often involving sarcasm, ambiguous viewpoints, and contextually nuanced arguments. Traditional NLP techniques, such as rule-based parsing, frequently struggle to handle these complexities, necessitating the development of more sophisticated approaches to effectively capture sentiment (Maynard and Funk, 2012).

Our study addresses the existing gap by focusing on the classification of political sentiment in Tamil tweets using a novel dataset and state-of-the-art machine learning models from the Hugging Face library. We systematically evaluate multiple transformer-based models and experiment with diverse hyperparameter configurations to achieve optimal performance. Our experiments demonstrate that the *"cardiffnlp/bert-base-multilingual-cased-sentiment-multilingual"* model achieves a training accuracy of 79%, with a precision of 80% and an F1 score of 80%. However, during validation, the model attains an accuracy of 33%, precision of 35%, and an F1 score of 34%. Despite the inherent challenges of sentiment analysis in Tamil, our results surpass those of other teams working on similar tasks, highlighting the potential of our approach. The findings of this study contribute to the development of more effective and reliable tools for political sentiment analysis in Tamil, addressing the needs of policymakers, researchers, and stakeholders interested in understanding Tamil public opinion. Our work paves the way for future advancements in sentiment analysis for low-resource languages, fostering deeper insights into political discourse in Tamil-speaking communities.

2 Related Works

Recent research on sentiment analysis of Tamil political comments using customized BERT models has focused on improving language comprehension and classification accuracy in Dravidian languages. TamilCogniBERT enhances Tamil text understanding through a pre-trained BERT framework with self-learning techniques (G et al., 2024). Task-specific pre-training and cross-lingual transfer learning have been shown to improve sentiment classification in Tamil-English code-mixed data (Gupta et al., 2021). Multi-task learning frameworks help tackle the issue of limited annotated data, enhancing sentiment and offensive language detection across Tamil, Malayalam, and Kannada (Hande et al., 2021). The DravidianCodeMix dataset, containing 60,000+ annotated comments, provides a strong foundation for model evaluation and training (Chakravarthi, 2022). Despite these advancements, challenges persist in handling diverse dialects and code-mixing, requiring further research to enhance model robustness.

3 Problem Description

Problem Statement. Sentiment analysis plays a crucial role in understanding public opinion, particularly in the political domain, where sentiments influence strategic decisions, policy-making, and public engagement. With the rise of social media platforms such as X (formerly Twitter), individuals now have a direct channel to express their political views (Chakravarthi et al., 2025). However, analyzing and classifying sentiments in Tamil tweets pose significant challenges due to the language's complex socio-cultural and linguistic characteristics.

The Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments Shared Task was conducted as an integral part of DravidianLangTech@NAACL 2025 (Chakravarthi et al., 2025). This task focused on the Political Multiclass

Sentiment Type	Train	Test
Opinionated	1361	153
Sarcastic	790	115
Neutral	637	84
Positive	575	69
Substantiated	412	52
Negative	406	51
None of the above	171	20

Table 1: Sentiment Distribution in Train and Test Sets

Sentiment Analysis of Tamil tweets, categorizing them into seven distinct sentiment classes: *Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral, and None of the Above*. Accurate classification requires not only linguistic proficiency but also an understanding of context, cultural nuances, and intent, adding substantial complexity to the task. Tamil, a Dravidian language with unique vocabulary and syntactic structures, poses challenges for conventional NLP techniques, especially in political discourse involving sarcasm and subjective expressions. Additionally, class imbalance in the dataset and the lack of robust pre-trained models for Tamil necessitate customized approaches for accurate sentiment classification.

Dataset. The dataset used for Tamil political sentiment analysis is divided into three subsets: training, validation, and testing. The **training set** consists of 4,352 Tamil political tweets, each labeled into one of seven sentiment classes: *Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral,* and *None of the Above*, serving as the primary source for model learning. The class distribution is added in Table 1. The **validation set** contains 544 labeled tweets and is used to fine-tune the model, ensuring generalization and preventing overfitting. Lastly, the **test set** comprises 544 unlabeled tweets, which are used for final evaluation by assessing the model's predictive performance. The dataset presents challenges such as class imbalance, with certain sentiment categories being underrepresented, making it difficult to achieve consistent accuracy across all classes. Despite these challenges, the dataset provides a valuable resource for developing and benchmarking sentiment analysis models tailored for Tamil political discourse.

4 System Description

Model. For sentiment analysis, we employ the *cardiffnlp/**b**ert-**b**ase-**m**ultilingual-**c**ased-**s**entiment-**m**ultilingual* model (BBMCSM, in short) (Antypas et al., 2022), which is a fine-tuned variant of the BERT base multilingual cased architecture. This model is specifically enhanced to classify sentiments in multilingual tweets, leveraging BERT's bidirectional processing capability to capture contextual meanings across various languages effectively. The model is designed for text classification tasks, particularly in analyzing sentiments expressed in social media content. The fine-tuning process utilizes the `TweetNLP` toolkit (Camacho-Collados et al., 2022), which is specialized for processing and analyzing tweet data. During evaluation on the test set, the model achieved an F1 score of 0.616 for both micro and macro measurements and an accuracy of 0.617, demonstrating a moderate level of reliability in sentiment detection for multilingual tweets. This model holds potential for automating sentiment analysis in social media, aiding in the understanding of public opinions and emotions across diverse linguistic contexts.

Implementation Details. The dataset used in this study consists of training, validation, and test sets, each serving a distinct purpose. The training set is utilized to optimize the model's weights through backpropagation, while the validation set is employed for hyperparameter tuning, ensuring the best configuration for generalization. The test set, which remains unseen during training, is used for the final evaluation of the model's performance on real-world data. For model training, we experimented with different hyperparameter configurations to achieve optimal performance. The final selected hyperparameters include an *input dimension* equal to the feature size of the training set, *number of classes* set to 7, and *hidden dimensions* of 1,536 and 786. We utilized a *batch size* of 32 and trained the model for *50 epochs* with a *learning rate* of 0.001 and a *dropout probability* of 0.3 to prevent overfitting. A fixed *random seed* of 42 was used to ensure reproducibility. Performance metrics such as Accuracy, Precision, Recall, and F1-Score were recorded to assess both training and validation outcomes. This systematic approach provides valuable insights for building efficient multilingual sentiment classification models with practical applications in political discourse analysis.

5 Experimental Findings

Training and Validation Results. The training and validation results demonstrated in Table 2

Table 2: Model Performance in Different Setups (Training and Validation Data)

Hidden dims	LR	dropout	T Acc	T Prec	T Rec	T F1	V Acc	V Prec	V Rec	V F1
1536, 786	0.001	0.3	0.7980	0.8040	0.7968	0.8002	0.3290	0.3429	0.3433	0.3391
1536, 786	0.001	0.4	0.7824	0.7924	0.7824	0.7850	0.3150	0.3285	0.3357	0.3262
1028, 786	0.001	0.3	0.6523	0.6711	0.6409	0.6527	0.3438	0.3496	0.339	0.3383
1028, 786	0.001	0.4	0.6250	0.6456	0.6123	0.6253	0.3334	0.3350	0.3243	0.3210
786, 256	0.001	0.3	0.3619	0.4052	0.2960	0.2582	0.3290	0.2654	0.2961	0.2382
786, 256	0.001	0.4	0.3557	0.3645	0.2834	0.2313	0.3107	0.2353	0.2675	0.2036
512, 256	0.001	0.3	0.3660	0.4086	0.3022	0.2692	0.2831	0.2284	0.2673	0.2254
512, 256	0.001	0.4	0.3500	0.3523	0.2934	0.2823	0.2934	0.2243	0.2723	0.2323

Table 3: Macro F1 Scores on Test Data

Submission	F1 Score (Macro)
Mean	0.2769
Median	0.2769
Our Result (Best)	**0.3187**
Our Rank	**4th**

shows the importance of hyperparameter optimization in achieving high performance for sentiment analysis tasks. Our experiments focused on fine-tuning a multilingual BERT model, with varying configurations of hidden dimensions, learning rates (LR), and dropout rates. The results highlight the influence of these hyperparameters on the model's ability to generalize across both training and validation sets. The results indicate that dropout plays a crucial role in preventing overfitting but can lead to underfitting when set too high. When the dropout rate is 0.3, the larger hidden dimensions (1536, 786) perform the best in both training and validation, achieving high accuracy, precision, recall, and F1-Score, with only a moderate drop in validation performance. However, increasing the dropout rate to 0.4 leads to a slight decrease in training performance and a more significant drop in validation results, especially for smaller hidden dimension configurations (1028, 786 and 512, 256). This suggests that while dropout helps regularize the model, a higher dropout rate can overly restrict the model's ability to learn from the data, particularly for smaller architectures. In the case of the 512, 256 configuration, the combination of smaller hidden dimensions and higher dropout results in poor training and validation performance, confirming the importance of selecting an appropriate model capacity for the task. Interestingly, the larger hidden dimensions maintain better generalization, as they are more robust to dropout, particularly in validation. This highlights the importance of balancing dropout and model size for optimal performance. Overall, a dropout rate of 0.3 is the most effective for achieving good generalization, particularly when using larger hidden dimensions, while higher dropout rates tend to hinder performance, especially for smaller models.

Test Results. Our customized BERT model also performed well on the test set, achieving an MF1 score of 0.3187, surpassing all other models. The average and median MF1 scores across all teams were 0.2769. This suggests that our approach, through hyperparameter optimization and improvement of the multilingual model, effectively captures sentiment patterns in Tamil political contexts. These results validate our approach and provide a foundation for further applications.

Overall, this study sets a benchmark for Tamil political sentiment analysis and opens avenues for future work, such as dataset expansion and exploring alternative architectures for multilingual sentiment analysis.

6 Conclusion

Sentiment analysis in low-resource languages presents unique challenges, and this work significantly contributes to Tamil political sentiment analysis. By developing a new annotated dataset and benchmarking transformer-based models, we demonstrated the feasibility of capturing subtle political sentiments in Tamil. Our fine-tuned multilingual BERT model achieved strong results, showcasing the effectiveness of NLP techniques and hyperparameter optimization in complex linguistic tasks. Despite challenges like class imbalance and nuanced sentiment expressions, this study provides a solid foundation for future research. The insights from this work can extend to other low-resource languages, advancing the goal of making NLP more inclusive across linguistic contexts.

Limitations

Despite our approach performing well, there are limitations to address. We were constrained by computational resources, preventing the use of larger, more complex models, which could improve accuracy by capturing deeper structures. Challenges such as class imbalance and ambiguous sentiments in the training and evaluation sets also impacted model performance. Additionally, while our fine-tuned multilingual model showed decent results, further domain-specific pretraining on Tamil data could enhance its understanding of political sentiment. Addressing these constraints in future work could lead to a more robust sentiment analysis framework.

Broader Impact

This work tackles sentiment analysis in low-resource languages, focusing on Tamil political discourse. By creating a new annotated dataset and experimenting with transformer-based models, we showcase the potential of multilingual BERT for capturing subtle political sentiments. Despite challenges like class imbalance and complex expressions, this study lays the foundation for future research in low-resource languages. Our findings contribute to making NLP more inclusive and adaptable across diverse linguistic and cultural contexts.

Acknowledgement

We express our sincere gratitude to Computational Intelligence and Operations Laboratory (CIOL) for their invaluable guidance, unwavering support, and continuous assistance throughout this journey. We are deeply appreciative of their efforts in organizing the CIOL Winter ML Bootcamp (Wasi et al., 2024), which provided an enriching learning environment and a strong foundation for collaborative research. The research mentoring and structured support offered by CIOL played a pivotal role in shaping this work, fostering innovation, and empowering participants to contribute meaningfully to the field of computational linguistics.

References

Mohd Zeeshan Ansari, M.B. Aziz, M.O. Siddiqui, H. Mehra, and K.P. Singh. 2020. Analysis of political sentiment orientations on twitter. *Procedia Computer Science*, 167:1821–1828.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Leonardo Neves, Vitor Silva, and Francesco Barbieri. 2022. Twitter Topic Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hila Becker, Mor Naaman, and Luis Gravano. 2021. Beyond trending topics: Real-world event identification on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):438–441.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, U.A.E. Association for Computational Linguistics.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11. ACM.

Bharathi R. Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Sajeetha Thavareesan, and Elizabeth Sherly, editors. 2023. *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria.

Bharathi Raja Chakravarthi. 2022. Dravidiancodemix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Parameswari Krishnamurthy, and Elizabeth Sherly, editors. 2021. *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Kyiv.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Parameswari Krishnamurthy, Elizabeth Sherly, and Sinnathamby Mahesan, editors. 2022. *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Dublin, Ireland.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Sajeetha Thavareesan, Elizabeth Sherly, Rajeswari Nadarajan, and Manikandan Ravikiran, editors. 2024. *Proceedings of the Fourth Workshop on Speech, Vision, and Language*

Technologies for Dravidian Languages. Association for Computational Linguistics, St. Julian's, Malta.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ashwinraj G, Sarfraz Hussain M, and Madhu Perkin T. 2024. Tamilcognibert: Enhancing tamil language comprehension using self learning. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–7.

Akshat Gupta, Sai Krishna Rallabandi, and Alan W. Black. 2021. Task-specific pre-training and cross lingual transfer for sentiment analysis in dravidian code-switched languages. page 73–79.

Adeep Hande, Siddhanth U Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages. *arXiv: Computational Complexity*.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, pages 538–541. AAAI Press.

Michael Mathioudakis and Nick Koudas. 2010. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD/PODS '10. ACM.

Diana Maynard and Adam Funk. 2012. *Automatic Detection of Political Opinions in Tweets*, page 88–99. Springer Berlin Heidelberg.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120, Jeju Island, Korea. Association for Computational Linguistics.

Azmine Toushik Wasi, MD Shakiqul Islam, Sheikh Ayatur Rahman, and Md Manjurul Ahsan. 2024. Ciol presnts winter ml bootcamp. 6 December, 2024 to 6 February, 2025.

Akatsuki-CIOL@DravidianLangTech 2025: Ensemble-Based Approach Using Pre-Trained Models for Fake News Detection in Dravidian Languages

Mahfuz Ahmed Anik[1], Md. Iqramul Hoque[1], Wahid Faisal[1], Azmine Toushik Wasi[1†], Md Manjurul Ahsan[2]
[1]Shahjalal University of Science and Technology, Sylhet, Bangladesh
[2]University of Oklahoma, Norman, OK 73019, USA
[†]**Correspondence:** azmine32@student.sust.edu

Abstract

The widespread spread of fake news on social media poses significant challenges, particularly for low-resource languages like Malayalam. The accessibility of social platforms accelerates misinformation, leading to societal polarization and poor decision-making. Detecting fake news in Malayalam is complex due to its linguistic diversity, code-mixing, and dialectal variations, compounded by the lack of large labeled datasets and tailored models. To address these, we developed a fine-tuned transformer-based model for binary and multiclass fake news detection. The binary classifier achieved a macro F1 score of 0.814, while the multiclass model, using multimodal embeddings, achieved a score of 0.1978. Our system ranked 14th and 11th in the shared task competition, highlighting the need for specialized techniques in underrepresented languages. Our full experimental codebase is publicly available at: ciol-researchlab/NAACL25-Akatsuki-Fake-News-Detection.

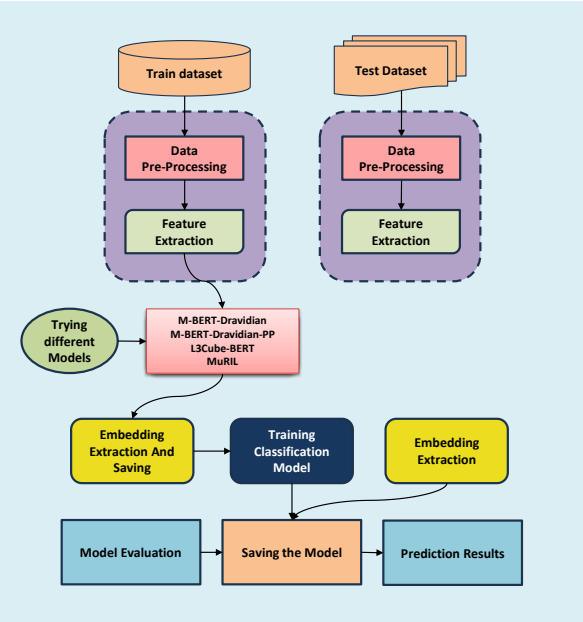


Figure 1: Model architecture, containing tokenizer, pre-trained model, classifier and other components

1 Introduction

Social media has transformed communication and information consumption, becoming a primary news source for many users worldwide. Platforms like Twitter, Facebook, and YouTube allow users to share and engage with information instantly, offering greater convenience, affordability, and timeliness compared to traditional news outlets (Kristian et al., 2024). These platforms solidify their role as preferred news mediums by enabling users to share, comment, and discuss news with their networks (Ku et al., 2019). However, this ease of sharing has also contributed to the widespread spread of fake news—misleading or false information designed to harm individuals, distort public opinion, or increase societal tensions (Fowler, Aug 22, 2022).

The spread of fake news on social media leads to emotional distress, societal polarization, and poor decision-making fueled by misinformation (De Paor and Heravi, 2020). Reports indicate that nearly 50% of Facebook referrals direct users to fake news sites (Pandey, 2018), while only 20% lead to reliable sources (Purcell et al., 2010). Furthermore, only 25% of individuals are confident in distinguishing real from fake news (Lyons et al., 2021), highlighting the need for scalable, automated solutions to address this growing issue.

Detecting fake news in low-resource languages like Malayalam is more challenging due to its linguistic diversity, including dialects, code-mixing, and idiomatic expressions (Thara and Poornachandran, 2021). The scarcity of structured datasets and pre-trained models for Malayalam compounds the problem (Elankath and Ramamirtham, 2023). Social media content, often containing code-mixed text in mixed scripts, poses further challenges for traditional NLP methods. While fake news detection has seen progress in high-resource languages

like English and Spanish, it remains underexplored for low-resource languages like Malayalam (Harris et al., 2024; Wang et al., 2024). Addressing misinformation in Dravidian languages is crucial, as they are spoken by millions in South India and Sri Lanka. The linguistic challenges, including code-switching and dialectal variations, necessitate tailored AI solutions for fairness and accuracy in fake news detection (Subramanian et al., 2025; Devika et al., 2024; Subramanian et al., 2023, 2024). Previous research has shown the effectiveness of ensemble-based models and feature fusion techniques in related tasks (Pillai and Arun, 2024).

In this study, we aim to bridge the gap in fake news detection for Malayalam by leveraging advanced NLP techniques and designing models that address its linguistic diversity and code-mixed nature, solving the first shared task of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025) at NAACL 2025. We use fine-tuned transformer-based architectures and hyperparameter optimization to develop robust solutions for binary and multiclass fake news classification tasks. Our approach tackles Malayalam's unique challenges, such as mixed scripts and dialectal variations, while ensuring scalability and effectiveness. This work provides valuable insights into the potential of transformer-based models for misinformation detection in low-resource languages, laying the foundation for future advancements in this area.

2 Problem Description

Problem Statement. The Fake News Detection in Dravidian Languages shared task focuses on addressing the critical issue of misinformation in low-resource languages, specifically in Tamil and Malayalam. The task is divided into two subtasks, each targeting a unique dimension of fake news detection:

Task 1, aims to classify social media posts from platforms like Twitter, Facebook, and YouTube into one of two categories: fake or original. This task operates at the comment or post level, challenging participants to build models that can effectively discern the authenticity of social media content.

Task 2, titled FakeDetect-Malayalam, targets the identification and classification of fake news within Malayalam-language news articles. Participants are tasked with categorizing news articles into one of five classes: False, Half True, Mostly False, Partly False, or Mostly True. This subtask emphasizes the nuanced detection of misinformation in a language with significant linguistic and cultural diversity.

Dataset. The datasets for the shared task are designed to support the development and evaluation of fake news detection models in Tamil and Malayalam. They are structured to address the unique requirements of the two subtasks.

The dataset for Task 1 consists of social media posts from platforms like Twitter, Facebook, and YouTube, labeled as either fake or original. This binary classification task aims to evaluate the authenticity of posts. The training set contains 3,257 labeled samples, while the validation set has 815 labeled samples for fine-tuning. The test set includes 1,019 unlabeled samples for model evaluation which is shown in 1.

The Task 2 dataset comprises Malayalam news articles categorized into five classes: False, Half True, Mostly False, Partly False, and Mostly True. This multiclass classification task focuses on detecting varying degrees of misinformation. The training dataset includes labeled articles, while the test dataset consists of unlabeled articles for evaluating participant systems.

Table 1: Dataset distribution for Task 1 and Task 2.

	Training	Development	Testing
Task 1	3,257	815	1,019
Task 2	1,615	285	200

3 System Description

3.1 Data Pre-processing

For both Task 1 and Task 2, we converted non-numerical labels into numerical representations for compatibility with machine learning models. Text sequences were tokenized using pre-trained tokenizers to retain domain-specific linguistic patterns and were truncated or padded to uniform lengths (512 tokens for Task 1 and 128 tokens for Task 2). Missing or invalid entries were removed to maintain data integrity. For Task 1, the label distribution was balanced, with 1,658 original and 1,599 fake samples in the train dataset. In contrast, Task 2 had significant class imbalance, with the following distribution: False (976), Mostly False (236), Half True (46), Partly False (129), and Mostly True (133). To address this, we applied a custom oversampling technique, replicating minority class samples to balance the dataset, reducing bias toward the

majority class and improving model generalization. Embeddings for both tasks were extracted from the [CLS] token of the BERT model to preserve contextual representations.

3.2 Models

For **Task 1**, we used "mdosama39/malayalam-bert-FakeNews-Dravidian" (M-BERT-Dravidian), a BERT-based model fine-tuned for Malayalam fake news detection, with 238 million parameters. It effectively captured contextual information for binary classification by extracting [CLS] token embeddings from the last hidden layer, which were then processed using a Multi-Layer Perceptron (MLP) classifier. The MLP had two hidden layers (786 and 512 dimensions) and a softmax output layer, refining the pre-trained features to distinguish between fake and original news.

For **Task 2**, we applied a multimodal embedding strategy, combining embeddings from two additional pre-trained models, "l3cube-pune/malayalam-bert" (L3Cube-BERT) and "Hate-speech-CNERG/hindi-abusive-MuRI" (MuRIL), to capture diverse linguistic and contextual features. This integration leveraged the strengths of multiple models to enable robust predictions for the complex multiclass classification task.

3.3 Implementation Details

We processed text sequences for both tasks using tokenizers from pre-trained models, truncating or padding inputs to 512 tokens for compatibility. For Task 1, we utilized the domain-specific tokenizer of M-BERT-Dravidian, extracting [CLS] token embeddings from the last hidden layer, which were fed into an MLP classifier with hidden dimensions of 786 and 512. For Task 2, we enhanced classification performance using a multimodel embedding strategy by combining embeddings from L3Cube-BERT and MuRIL. These concatenated embeddings were processed through the same MLP architecture, leveraging the linguistic and contextual diversity of the combined models.

Both tasks employed a batch size of 16, the Adam optimizer (learning rate: 0.0001, betas: 0.9, 0.999), and linear learning rate scheduling. Figure 1 illustrates the overall system architecture. A dropout rate of 20% was applied for Task 1 to mitigate overfitting, while no dropout was used for Task 2 to maintain the integrity of combined embeddings. Training and evaluation pipelines were implemented in a GPU-enabled environment using PyTorch (v2.0.0) and Hugging Face Transformers (v4.35.0), ensuring efficient computation. Table 2 summarizes the hyperparameters, hidden dimensions, batch sizes, and performance metrics for all models.

4 Experimental Findings

4.1 Training and Validation Results

For **Task 1**, our best-performing model, M-BERT-Dravidian, achieved a training F1 score of 0.9794 and a validation F1 score of 0.8304, as shown in Table 2. These results demonstrate strong performance with effective generalization from the training data to the validation set. The minimal gap between training and validation scores highlights the robustness of the model, indicating no significant overfitting during training.

For **Task 2**, the best validation F1 score was achieved by the ensemble of MuRIL and L3Cube-BERT, which obtained a training F1 score of 0.9890 and a validation F1 score of 0.4115. This result highlights the ensemble's ability to handle the linguistic diversity and class imbalance challenges inherent to Task 2. The slight performance improvement over other models indicates the potential benefits of combining features from multiple pre-trained models for low-resource languages.

4.2 Test Results

As shown in Table 3, our model achieved a macro F1 score of 0.814 for Task 1 and 0.1978 for Task 2 on the test dataset. For Task 1, our score is close to the highest score of 0.898, and above the mean (0.7805) and median (0.832), demonstrating strong performance in binary classification. The minimum score of 0.334 further highlights the model's effectiveness. In contrast, Task 2 presented greater challenges, with a score of 0.1978, below the mean (0.3244) and median (0.2593), and far behind the top-performing system's score of 0.6283. These results emphasize the model's robustness in Task 1 and reveal the complexities of multiclass classification in low-resource, code-mixed settings.

4.3 Ablation Studies

We evaluated several models for Task 1 and Task 2, as shown in Table 2. For Task 1, M-BERT-Dravidian achieved the best validation F1 score of 0.8304, showing strong generalization with balanced precision and recall. The combined model using embeddings from M-BERT-Dravidian,

Table 2: Hyperparameter Settings and Performance Metrics for Task 1 and Task 2 Train and Validation Dataset.

Task	Model	Max Length	Batch Size	Hidden Dim	LR	Dropout	Train Acc	Train F1	Val Acc	Val F1
Task 1	M-BERT-Dravidian	512	16	[786, 512]	0.0001	0.2	0.9794	0.9794	0.8307	0.8304
Task 1	M-BERT-Dravidian, L3Cube-BERT, MuRIL	512, 512, 512	8	[786, 512]	0.0001	0.0	0.9942	0.9942	0.827	0.827
Task 1	L3Cube-BERT	786	16	[786, 512]	0.0001	0.3	0.9975	0.9975	0.8258	0.8255
Task 2	M-BERT-Dravidian-PP	512	16	[768, 512]	0.0001	0.35	0.7771	0.7720	0.5474	0.3916
Task 2	M-BERT-Dravidian, L3Cube-BERT	512	16	[768, 512]	0.0001	0.35	0.8758	0.8743	0.5921	0.4097
Task 2	MuRIL, L3Cube-BERT	512, 512	8	[786, 512]	0.0001	0.5	0.9890	0.9890	0.6351	0.4115

Table 3: F1 Score (Macro) on Test Dataset

Macro F1	Maximum	Minimum	Mean	Median	Our Score
Task 1	0.898	0.334	0.7805	0.832	0.814
Task 2	0.6283	0.1667	0.3244	0.2593	0.1978

L3Cube-BERT, and MuRIL had a slightly lower validation F1 score of 0.827, with a high training F1 score of 0.9942, indicating overfitting. L3Cube-BERT achieved a validation F1 score of 0.8255, demonstrating good training performance but less robustness. For Task 2, traditional methods like Bag of Words and TF-IDF (Dai et al., 2024; Deo et al., 2024) were less effective in complex multiclass tasks. Among individual models, M-BERT-Dravidian-PP achieved the best validation F1 score of 0.3916 and a training F1 score of 0.7720. Combining models improved performance, with the L3Cube-BERT and MuRIL ensemble achieving a validation F1 score of 0.4115 and a training F1 score of 0.9890. Another combination, M-BERT-Dravidian-PP and L3Cube-BERT, achieved a validation F1 score of 0.4068, despite strong training performance (F1: 0.9783). These results highlight the effectiveness of model ensembles for multiclass classification in Task 2.

For Task 1, the best validation F1 score of 0.8304 was achieved at epoch 46. Early in training, the model showed overfitting with a high training F1 score of 0.9874 at epoch 1, while the validation F1 score lagged at 0.8110. However, generalization improved, reaching 0.8232 at epoch 38 before peaking at epoch 46. For Task 2, M-BERT-Dravidian-PP achieved its best validation F1 score of 0.3916 at epoch 19, with a training F1 of 0.7720. Combined models performed better, with the L3Cube-BERT and MuRIL ensemble achieving the highest validation F1 score of 0.4115 at epoch 27. Another combination, M-BERT-Dravidian-PP and L3Cube-BERT, peaked at epoch 31 with a validation F1 score of 0.4068, but was slightly less robust than the top ensemble. These results emphasize the importance of training duration and model synergy for optimal performance.

5 Discussion

This study explored fake news detection in Malayalam through binary and multiclass classification. In Task 1, our fine-tuned transformer model achieved a macro F1 score of 0.814, effectively distinguishing fake from authentic posts with limited labeled data. Task 2 was more challenging, with a macro F1 score of 0.1978, requiring classification into nuanced categories like "Half True" and "Partly False." This task highlighted the complexities of dialectal variations, class imbalance, and code-mixed text in Malayalam. The performance gap between binary and multiclass classification shows the need for larger datasets, enhanced augmentation, and class-aware loss functions. Our study demonstrates the potential of transformer models and multimodal embeddings for Malayalam's linguistic diversity.

6 Conclusion

This study investigated fake news detection in Malayalam through binary and multiclass classification tasks. The binary classifier achieved a strong macro F1 score of 0.814, showcasing the effectiveness of transformer-based models in identifying misinformation in social media posts. However, the lower performance in multiclass classification (macro F1 score of 0.1978) underscores the challenges of categorizing nuanced misinformation, which requires deeper contextual understanding and tailored strategies. These findings highlight the importance of diverse training data, robust preprocessing, and context-aware approaches to address linguistic complexities such as code-mixing and dialect variations. Future research should aim to advance fake news detection in low-resource languages for greater effectiveness and inclusivity.

Limitations

Despite promising results, limitations persist. The small dataset size, especially for multiclass classification, restricts the model's ability to capture Malayalam's dialects and script variations. Although oversampling addressed class imbalance, it introduced risks of bias and overfitting (Gosain and Sardana, 2017). Differentiating similar misinformation categories remains challenging, requiring architectures capable of finer semantic distinctions. Dependence on pre-trained models risks propagating biases from their training data. While ensemble methods boosted performance, they increased computational complexity, limiting scalability. Lastly, curated datasets may not fully reflect real-world social media complexities, emphasizing the need for diverse data and adversarial training to enhance generalization (De Paor and Heravi, 2020).

Broader Impact Statement

Developing robust fake news detection systems for low-resource languages like Malayalam can significantly curb misinformation in underrepresented communities, fostering informed decision-making and social harmony. Addressing linguistic challenges like code-mixing and dialectal diversity contributes to inclusive AI solutions, bridging resource gaps in NLP. These advancements promote media literacy and trust in digital platforms, mitigating societal polarization and ensuring equitable access to reliable information.

Acknowledgement

We express our sincere gratitude to Computational Intelligence and Operations Laboratory (CIOL) for their invaluable guidance, unwavering support, and continuous assistance throughout this journey. We are deeply appreciative of their efforts in organizing the CIOL Winter ML Bootcamp (Wasi et al., 2024), which provided an enriching learning environment and a strong foundation for collaborative research. The research mentoring and structured support offered by CIOL played a pivotal role in shaping this work, fostering innovation, and empowering participants to contribute meaningfully to the field of computational linguistics.

References

Shuying Dai, Keqin Li, Zhuolun Luo, Peng Zhao, Bo Hong, Armando Zhu, and Jiabei Liu. 2024. AI-based nlp section discusses the application and effect of bag-of-words models and tf-idf in nlp tasks. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 5(1):13–21.

Saoirse De Paor and Bahareh Heravi. 2020. Information literacy and fake news: How the field of librarianship can help combat the epidemic of fake news. *The Journal of Academic Librarianship*, 46(5):102218.

Saoirse De Paor and Bahareh Heravi. 2020. Information literacy and fake news: How the field of librarianship can help combat the epidemic of fake news. *The Journal of Academic Librarianship*, 46(5):102218.

Tula Kanta Deo, Rajesh Keshavrao Deshmukh, and Gajendra Sharma. 2024. Comparative study among term frequency-inverse document frequency and count vectorizer towards k nearest neighbor and decision tree classifiers for text dataset. *Nepal Journal of Multidisciplinary Research*, 7(2):1–11.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Syam Mohan Elankath and Sunitha Ramamirtham. 2023. Sentiment analysis of malayalam tweets using bidirectional encoder representations from transformers: a study. *Indonesian Journal of Electrical Engineering and Computer Science*, 29(3):1817–1826.

Gary Fowler. Aug 22, 2022. Council Post: Fake News, Its Impact And How Tech Can Combat Misinformation — forbes.com. [Accessed 26-01-2025].

Anjana Gosain and Saanchi Sardana. 2017. Handling class imbalance problem using oversampling techniques: A review. In *2017 international conference on advances in computing, communications and informatics (ICACCI)*, pages 79–85. IEEE.

Sheetal Harris, Hassan Jalil Hadi, Naveed Ahmad, and Mohammed Ali Alshara. 2024. Fake news detection revisited: An extensive review of theoretical frameworks, dataset assessments, model constraints, and forward-looking research agendas. *Technologies*, 12(11):222.

Natalia Kristian, Dana Indra Sensuse, and Sofian Lusa. 2024. The role of social media functions in enhancing knowledge sharing with user engagement and information quality. *Jurnal Indonesia Sosial Teknologi*, 5(10).

Kelly YL Ku, Qiuyi Kong, Yunya Song, Lipeng Deng, Yi Kang, and Aihua Hu. 2019. What predicts adolescents' critical thinking about real-life news? the roles of social media news consumption and news media literacy. *Thinking Skills and Creativity*, 33:100570.

Benjamin A Lyons, Jacob M Montgomery, Andrew M Guess, Brendan Nyhan, and Jason Reifler. 2021. Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, 118(23):e2019527118.

Neha Pandey. 2018. Fake news: A manufactured deception, distortion and disinformation is the new challenge to digital literacy. *Journal of Content, Community and Communication*, 4(8):15–21.

Aditya R Pillai and Biri Arun. 2024. A feature fusion and detection approach using deep learning for sentimental analysis and offensive text detection from code-mix malayalam language. *Biomedical Signal Processing and Control*, 89:105763.

Kristen Purcell, Lee Rainie, Amy Mitchell, Tom Rosenstiel, and Kenny Olmstead. 2010. Understanding the participatory news consumer. *Pew Internet and American Life Project*, 1:19–21.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

S Thara and Prabaharan Poornachandran. 2021. Transformer based language identification for malayalam-english code-mixed text. *IEEE Access*, 9:118837–118850.

Xinyu Wang, Wenbo Zhang, and Sarah Rajtmajer. 2024. Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey. *arXiv preprint arXiv:2410.18390*.

Azmine Toushik Wasi, MD Shakiqul Islam, Sheikh Ayatur Rahman, and Md Manjurul Ahsan. 2024. Ciol presnts winter ml bootcamp. 6 December, 2024 to 6 February, 2025.

A Appendix

A.1 Error Analysis

For **Task 1**, fig 2 shows that 433 Fake samples were correctly labeled, with 74 misclassified as Original. Conversely, 443 Original samples were correct, but 69 were wrongly labeled as Fake. For **Task 2**, fig 3 142 out of 149 False samples were accurately identified. However, the model struggled with more nuanced classes: 16 out of 24 Half True were misclassified, 78.57% of Partly False were incorrectly labeled, and Only 41.27% of Mostly False samples were correct. This bias toward False likely stems from its dominance (three-fourths) in the development data. Sample prediction errors and actual labels are included in table 4. Incorrect predictions largely arise from limited context in short or slang-laden Malayalam text, data imbalance where minority classes (e.g., "Mostly False") are often misclassified, and subtle semantic overlaps between similar labels (e.g., "Fake" vs. "Original"). Such subtleties are challenging for the model to detect without sufficient training examples or language-specific fine-tuning, highlighting the need for data augmentation, balanced class distribution, and more extensive contextual cues to improve classification accuracy.

Table 4: Incorrect Predictions in Text Classification

Task	Text Sample	Predicted	Actual
Task-1	Sample:പരാജയം	Fake	original
Task-1	Sample: ചുവന്ന ട്രൗസർ ഇട്ടാൽ കൊറോണ വരില്ല എന്ന് അറിയില്ലേ ഗമ്മേ	original	Fake
Task-1	Sample: താബിലീഗ് ക്കാർ പിന്നെ അവർ ആവർ ത്തിച്ചി രു ന്നങ്കിൽ നമുക് പറയാൻ മായിരുന്നു. എന്നഈ ഗുണ്ടി മേളം നിയന്ത്രി ക്കാൻ കഴിഞ്ഞില	Fake	original
Task-2	Sample: മഞ്ഞ് ഉരുകുന്നില്ല, കറുത്തിരുണ്ടു പൊള്ളുന്നു	Mostly False	False
Task-2	Sample: ബഫല്ലോ മഞ്ഞുവീഴ്ചയെത്തുടർന്ന് കുടുങ്ങിയ ഒരു കാർ ഫോട്ടോ കാണിക്കുന്നു	Half True	False
Task-2	Sample: ബിബിൻ ജോർജിനെ കോളേജിൽ നിന്നും അപമാനിച്ച് ഇറക്കിവിട്ട സംഭവം ചർച്ചയാകാത്തത് ഇറക്കി വിട്ടയാൾ മുസ്ലിമായതിനാൽ.	False	Mostly False

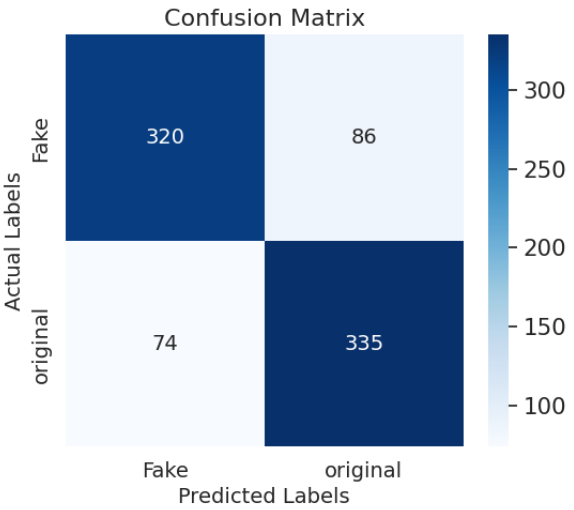


Figure 2: Confusion Matrix of task 1

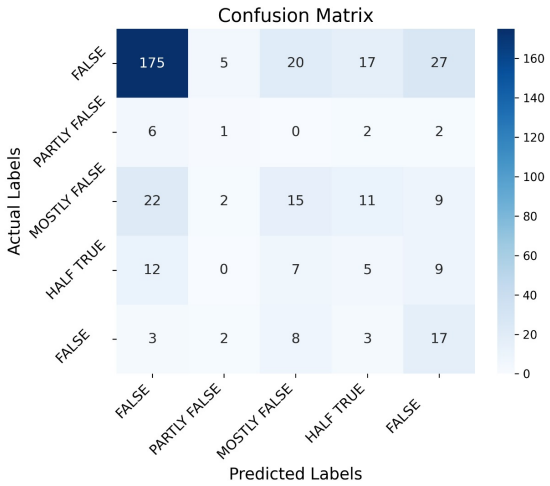


Figure 3: Confusion Matrix of task 2

RMKMavericks@DravidianLangTech 2025: Tackling Abusive Tamil and Malayalam Text Targeting Women: A Linguistic Approach

Sandra Johnson
R.M.K. Engineering College
Tiruvallur
sjn.ad@rmkec.ac.in

Boomika E
R.M.K. Engineering College
Tiruvallur
boom22011.ad@rmkec.ac.in

Lahari P
R.M.K. Engineering College
Tiruvallur
laha22024.ad@rmkec.ac.in

Abstract

Social media abuse of women is a widespread problem, especially in regional languages like Tamil and Malayalam, where there are few tools for automated identification. The use of machine learning methods to detect abusive messages in several languages is examined in this work. An external dataset was used to train a Support Vector Machine (SVM) model for Tamil, which produced an F1 score of 0.6196. Using the given dataset, a Multinomial Naive Bayes (MNB) model was trained for Malayalam, obtaining an F1 score of 0.6484. Both models processed and analyzed textual input efficiently by using TF-IDF vectorization for feature extraction. This method shows the ability to solve the linguistic diversity and complexity of abusive language identification by utilizing language-specific datasets and customized algorithms. The results highlight how crucial it is to use focused machine learning techniques to make online spaces safer for women, especially when speaking minority languages.

1 Introduction

Social media pervasiveness has changed how people interact, but it has also contributed to the growth of abusive content, which mostly targets women and other vulnerable groups. Since this type of online harassment has detrimental effects on mental and emotional health, it is critical to create efficient detection and moderation systems. Because of their distinct linguistic traits and the dearth of techniques for detecting abusive content, regional languages like Tamil and Malayalam make the issue much more difficult. By using machine learning methods designed especially for the Tamil and Malayalam languages, this study seeks to close that gap. Specifically, Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) models are used to categorize offensive text in various languages.

https://github.com/Boomika2005/DravidianLangTech-Abusive-detection

While MNB is used for Malayalam abuse detection, SVM, which is renowned for its capacity to handle complicated and non-linear data, is used to Tamil. TF-IDF vectorization, which captures the essential characteristics of text data for classification, supports both models. By utilizing these strategies, this study aims to offer a strong system for identifying and filtering offensive language in local settings, making online environments safer for women.

2 Related Work

To Recent research has focused extensively on offensive language detection and sentiment analysis in Dravidian languages, addressing the challenges posed by the use of code-mixed and multilingual text on social media platforms. Machine learning models built with monolingual datasets are often inadequate for identifying abusive language or analyzing sentiments from code-mixed languages, which blend multiple languages such as Tamil, Malayalam, and English. Numerous research have advanced this subject by creating models and datasets especially suited for Dravidian languages. Datasets for hate speech analysis and objectionable language identification were published by Chakravarthi et al. (2020), Hande et al. (2021), and Mandl et al. (2020). These datasets are now vital tools for scholars developing Dravidian language models. In order to extract contextual characteristics from Tamil, Malayalam, and Kannada text, Saumya et al. (2021) used CNN and Bi-LSTM models. High F1-scores of 0.7895 for Tamil and 0.9603 for Malayalam were attained by their effort. In a similar vein, Yasaswini et al. (2021) used the ULMFiT model and obtained F1-scores of 0.7895 for Tamil and 0.9603 for Malayalam. For Dravidian code-mixed languages, Kedia and Nandy (2021) suggested transformer-based models, such as BERT and RoBERTa, which achieved weighted average

F1-scores of 0.72, 0.77, and 0.97 for the datasets pertaining to Kannada-English, Tamil-English, and Malayalam-English, respectively. The application of transfer learning techniques and cross-lingual contextual word embeddings has also been studied. Multinomial Naive Bayes, SVM, and Random Forest were tested by Ranasinghe et al. (2020), who obtained an F1-score of 0.89 for code-mixed Malayalam. Furthermore, Sai and Sharma (2021) have used an ensemble of multilingual transformer networks, such as XLMRoBERTa, for the identification of objectionable speech in Tamil, Malayalam, and Kannada. The investigation of multimodal datasets is still in its infancy, despite notable advancements in text-based datasets and algorithms for sentiment analysis and abusive language identification. A more thorough method for examining abusive language in Dravidian languages may be offered by combining textual data with audio and visual data.

3 Preprocessing and Data Preparation

Preprocessing and data preparation are essential steps to prepare the datasets for machine learning models. For this work, datasets were provided by the organizers, with separate training and testing datasets for both Tamil and Malayalam. Additionally, an external dataset was used for Tamil to enhance the diversity and performance of the model. Preprocessing steps were carefully designed to handle the challenges posed by code-mixed text, ensuring the preservation of linguistic nuances and optimal input for feature extraction techniques.

3.1 Data Refinement

The first step in preprocessing involved refining the text data by cleaning and normalizing it. All text was converted to lowercase to ensure uniformity and eliminate issues related to case sensitivity. Special characters, numbers, and punctuation marks were removed unless they contributed meaningful context to the text. Tokenization was applied to break the text into smaller units, such as words or subwords, for easier processing. For code-mixed data, specific considerations were made to retain the integrity of the mixed-language structure. Stopwords that did not contribute to the task were removed selectively, and whitespace inconsistencies were corrected.

3.2 Feature Extraction

Following data refinement, textual data was transformed into numerical representations for the machine learning models using feature extraction techniques:

Bag of Words (BOW): Text was represented as a vector of word frequencies using the Bag of Words (BOW) technique. By capturing word occurrences across the dataset, this method enabled the model to use the frequency of particular phrases to make predictions.

TF-IDF (Term Frequency-Inverse Document Frequency): Words were weighed according to their significance in the dataset using TF-IDF. This approach emphasized uncommon but important keywords while lessening the effect of often recurring ones. These feature extraction methods preserved the significant connections between words and their context in code-mixed languages while guaranteeing that the text data was converted into a machine learning-ready format.

4 Methodology

4.1 Model Selection

In order to efficiently categorize code-mixed Tamil and Malayalam texts, various machine learning models were investigated in this work. Support Vector Machines (SVM) were chosen for Tamil because of its ability to handle high-dimensional, complicated data, whereas Multinomial Naive Bayes was picked for Malayalam because of its ease of use and efficacy when processing categorical data.

4.2 Training of Models

In order to develop models that could successfully categorize code-mixed Dravidian languages, Tamil, and Malayalam, the training phase was essential. The organizers' datasets were pre-processed using methods like Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) to provide structured numerical representations. By using these techniques, the textual input was converted into valuable characteristics that could be used to train the model. Support Vector Machines (SVM) was chosen as the classification technique for Tamil because of its resilience in text classification tasks and its ability to handle high-dimensional feature spaces. To maximize performance, a grid search method was used for hyperparameter optimization during the training phase. Figure 1 (Tamil Grid Search Results: Accuracy by C and Kernel)

showed the accuracy trends that resulted from the grid search's evaluation of combinations of the regularization parameter (C) and kernel types. To have the highest prediction accuracy for the Tamil dataset, this tuning procedure was essential.

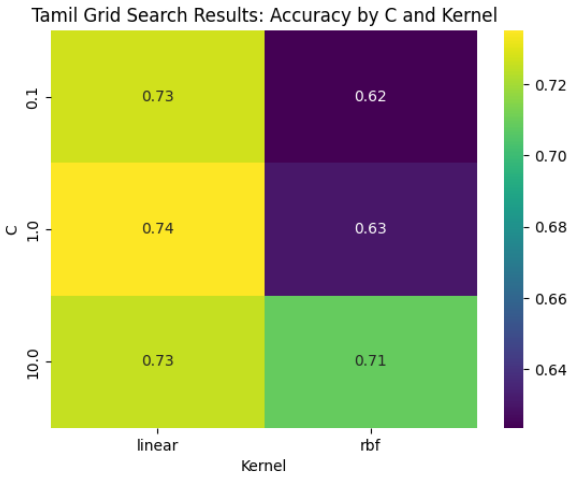


Figure 1: Tamil Grid Search Results: Accuracy by C and Kernel

Multinomial Naive Bayes was selected for Malayalam because of its performance in text classification tasks that use features based on word frequency. It was especially appropriate for this dataset due to its ease of use and computational effectiveness. In contrast to Tamil, which required hyperparameter adjustment, the Naive Bayes model was trained straight from the processed dataset without the need for further optimization. The goal of training these models was to take use of the distinctive features of the individual datasets. To guarantee a well-rounded predictive capacity, the focus was on optimizing performance indicators like accuracy and F1-score throughout the training phase. The foundation for the following phases of performance analysis and assessment was established by this methodical methodology.

4.3 Model Performance Evaluation Metrics

Several measures were used to evaluate the models' performance and determine their capabilities. The models' efficacy was assessed using Accuracy, Precision, Recall, F1-score, and Area Under the Curve (AUC). To illustrate the relationship between the anticipated and real labels, confusion matrices were also produced (Figure 2).

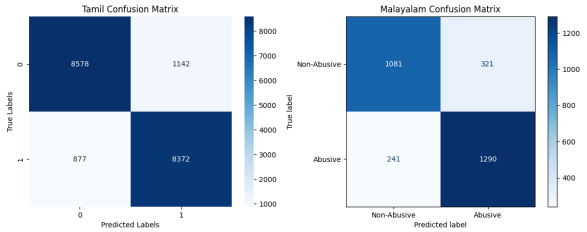


Figure 2: Confusion matrix Predicted and True labels for Tamil & Malayalam

4.4 Area Under the Curve (AUC) and ROC Curve

The models' overall performance was assessed using the AUC-ROC curve, which is displayed in Figure 3. This curve clearly illustrates how well the model performs in various settings by plotting the true positive rate (Recall) versus the false positive rate (1-Specificity) at various thresholds. The overall indicator of model performance is the AUC value; a higher AUC denotes a model that performs better overall.

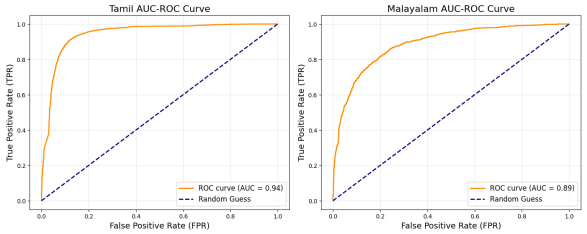


Figure 3: AUC-ROC curve Tamil & Malayalam.

4.5 Precision-Recall Curve

The balance between precision and recall was further assessed using the Precision-Recall curve, which is shown in Figure 4. This curve provides insight into the model's performance in detecting positive examples across various thresholds. Both the Tamil and Malayalam datasets accuracy and recall trade-offs are displayed in the graphic.

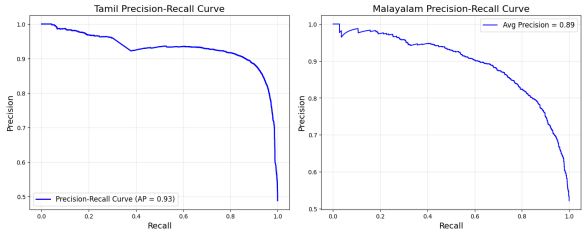


Figure 4: Precision-Recall Curve for Tamil and Malayalam datasets.

5 Result and Findings

Several performance indicators were used to assess the models that were trained for Malayalam and Tamil. A comprehensive understanding of model performance was made possible by the classification report for both languages, which included information on the precision, recall, and F1-score for each class. In handling code-mixed Dravidian languages, these findings showed how well the chosen algorithms—SVM for Tamil and Multinomial Naive Bayes for Malayalam—balanced accuracy and efficiency.

Metrics	Tamil	Malayalam
Accuracy	0.62	0.65
Precision	0.62	0.65
Recall	0.62	0.65
Macro F1 Score	0.6196	0.6484

Table 1: Performance metrics for Tamil and Malayalam tasks.

6 Conclusion

The difficulties of dealing with code-mixed Dravidian languages were addressed in this work by developing models for Tamil and Malayalam. Multinomial Naive Bayes was utilized for Malayalam, and Support Vector Machines (SVM) for Tamil. Both models showed excellent performance in categorizing abusive language and feelings through efficient preprocessing, feature extraction, and model training. Balanced accuracy, recall, and F1-scores were found in the evaluation measures, demonstrating the models' capacity to manage the complexity of code-mixed data. These results lay the groundwork for future studies and advancements in the field of natural language processing (NLP) for Dravidian languages.

7 Limitations

Even while this study produced encouraging results, it must be noted that it has significant limitations. A significant obstacle was managing the extremely informal and unstructured character of code-mixed Dravidian languages, which frequently have intricate linguistic variances. Due to the small dataset size, especially for Malayalam, the model's generalizability may have been impacted. Using TF-IDF and Bag of Words (BOW) representations alone could also leave out important contextual and semantic links between words. Although they might improve performance even further, advanced deep learning models like transformer-based topologies were not investigated in this work because of computing limitations. Adding contextual embeddings, broadening datasets, and enhancing generalization across dialects and variances might be the main goals of future research.

References

Judith Jeyafreeda Andrew. 2021. Judithjeyafreedaandrew@dravidianlangtech-eacl2021: Offensive language detection for dravidian code-mixed youtube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, page 169–174.

Noman Ashraf, Arkaitz Zubiaga, and Alexander Gelbukh. 2021. Abusive language detection in youtube comments leveraging replies as conversational context. *PeerJ Computer Science*, 7:e742.

Shubhankar Barman and Mithun Das. 2023. hatealert@dravidianlangtech: Multimodal abusive language detection and sentiment analysis in dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*.

Darrell Davis, Ranjith Murali, and Ramesh Babu. 2020. Abusive language detection and characterization of twitter behavior. *arXiv preprint*, arXiv:2009.14261.

Tariq Kanan, Ahmed Aldaaja, and Bilal Hawashin. 2020. Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in arabic social media contents. *Journal of Internet Technology*, 21(5):1409–1421.

Simran Kaur, Sukhpreet Singh, and Sunil Kaushal. 2021. Abusive content detection in online user-generated data: A survey. *Procedia Computer Science*, 189:274–281.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153.

SN Prasanth, R Aswin Raj, P Adhithan, B Premjith, and Soman Kp. 2022. Centamil@dravidianlangtechacl2022: Abusive comment detection in tamil using tf-idf and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, page 70–74.

B. Premjith, G. Jyothish Lal, V. Sowmya, B.R. Chakravarthi, R. Natarajan, K. Nandhini, A. Murugappan, B. Bharathi, M. Kaushik, S.N. Prasanth, R.A. Raj, and S.V. Vijai Simmon. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *DravidianLangTech 2023 - 3rd Workshop on Speech and Language Technologies for Dravidian Languages*. RANLP 2023.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, page 292–298. Association for Computational Linguistics.

Sudarshan Rajamanickam, Prashant Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. *arXiv preprint*, arXiv:2005.14028.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Association for Computational Linguistics*.

RMKMavericks@DravidianLangTech 2025: Emotion Mining in Tamil and Tulu Code-Mixed Text: Challenges and Insights

Gladiss Merlin N.R
R.M.K. Engineering College
Tiruvallur
nrg.ad@rmkec.ac.in

Boomika E
R.M.K. Engineering College
Tiruvallur
boom22011.ad@rmkec.ac.in

Lahari P
R.M.K. Engineering College
Tiruvallur
laha22024.ad@rmkec.ac.in

Abstract

Sentiment analysis in code-mixed social media comments written in Tamil and Tulu presents unique challenges due to grammatical inconsistencies, code-switching, and the use of non-native scripts. To address these complexities, we employ pre-processing techniques for text cleaning and evaluate machine learning models tailored for sentiment detection. Traditional machine learning methods combined with feature extraction strategies, such as TF-IDF, are utilized. While logistic regression demonstrated reasonable performance on the Tamil dataset, achieving a macro F1 score of 0.44, support vector machines (SVM) outperformed logistic regression on the Tulu dataset with a macro F1 score of 0.54. These results demonstrate the effectiveness of traditional approaches, particularly SVM, in handling low-resource, multilingual data, while also highlighting the need for further refinement to improve performance across underrepresented sentiment classes.

1 Introduction

The growing use of social media platforms has led to an abundance of user-generated content, often expressed in code-mixed languages. Tamil and Tulu, two Dravidian languages, frequently appear in such code-mixed forms, blending with English and other languages. These multilingual and code-mixed texts introduce significant challenges for sentiment analysis due to their informal grammar, irregular structures, and non-standard scripts.

Sentiment analysis aims to identify and classify subjective opinions or emotions expressed in text. While considerable progress has been made in analyzing texts in resource-rich languages, low-resource languages like Tamil and Tulu remain underexplored. Existing tools and models, primarily designed for monolingual texts, struggle to perform effectively on code-mixed data, necessitating novel approaches tailored to these contexts.

In this work, we investigate traditional machine learning methods for sentiment analysis on Tamil-English and Tulu-English code-mixed datasets. We leverage pre-processing techniques to clean and normalize the data, employ TF-IDF (Term Frequency-Inverse Document Frequency) for feature extraction, and evaluate multiple machine learning classifiers. By optimizing hyperparameters and focusing on feature engineering, we aim to improve sentiment classification performance for these low-resource, multilingual datasets. The results underscore the importance of adapting traditional techniques to the unique challenges posed by code-mixed text data.

2 Related Work

Sentiment analysis has been a prominent area of research for several decades, focusing on identifying and classifying emotions, opinions, and attitudes expressed in text. Traditional approaches to sentiment analysis can be categorized into lexicon-based, machine learning-based, and hybrid methods. Lexicon-based approaches rely on predefined sentiment dictionaries to determine the polarity of text. Machine learning-based methods, often using supervised algorithms, leverage labeled datasets to train models for sentiment classification. Hybrid methods combine the strengths of both approaches, aiming to improve accuracy across diverse datasets.

Research in sentiment analysis for low-resource languages, such as Tamil and Tulu, has gained attention more recently. Early studies predominantly relied on rule-based systems or basic machine learning techniques, often limited by the availability of annotated datasets and language-specific tools. For example, Thavareesan and Mahesan (2020a) explored machine learning techniques for Tamil text sentiment analysis using feature representations like word embeddings and TF-IDF.

https://github.com/Boomika2005/RMKMavericks-Sentiment-analysis

Sentiment analysis of code-mixed text introduces additional complexities due to frequent switching between languages, irregular grammar, and the use of non-native scripts.

Recent studies have also examined the use of deep learning models, such as recurrent neural networks and transformer architectures, for low-resource and code-mixed languages. However, these methods typically require substantial computational resources and large annotated datasets, which are not always available for Tamil and Tulu.

Building on this body of work, our research focuses on traditional machine learning models optimized with feature extraction techniques, such as TF-IDF, and hyperparameter tuning. By leveraging these methods, we aim to address the unique challenges posed by Tamil-English and Tulu-English code-mixed datasets.

3 Task details

Sentiment analysis refers to the process of determining the emotional tone or subjective opinions expressed in a given piece of text. This area of research has gained significant attention over the past two decades, both in academic and industry settings. With the rise of social media, there is an increasing demand for systems that can analyze sentiment in social media posts, which are often written in code-mixed languages, particularly in Dravidian languages. Code-mixing, the practice of blending multiple languages in a single sentence or passage, is common in multilingual communities, and these texts may sometimes be written in non-native scripts. Traditional systems trained on monolingual datasets struggle with code-mixed text due to the complexities of language switching and its varying impact on grammar, syntax, and vocabulary.

The objective of this task is to determine the sentiment polarity of code-mixed comments/posts in Tamil-English and Tulu-English, collected from social media platforms. While each comment/post may consist of more than one sentence, the average sentence length in this dataset is short. Each comment/post is labeled with a sentiment polarity, either positive, negative, or neutral. The dataset also includes class imbalance, reflecting the real-world challenges encountered in sentiment analysis applications.

This task encourages further exploration into how sentiment is expressed in code-mixed texts, particularly in the context of social media communications.

4 Methodology

Our approach for sentiment analysis in the shared task involved implementing two traditional machine learning models and performing various data processing steps to handle the challenges of code-mixed text. We began by importing essential libraries such as Pandas, NumPy, and scikit-learn for tasks like data loading, cleaning, tokenization, vectorization, and modeling.

First, we loaded the training and validation datasets using Pandas, which contained code-mixed Tamil-English and Tulu-English comments/posts. We then cleaned the data by removing unnecessary punctuation and converting the text to lowercase to ensure consistency. This preprocessing step helped improve the models' ability to detect sentiment accurately.

After cleaning the text, we applied the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer to transform the text data into a numerical format that can be used by machine learning models. We selected unigrams and bigrams for tokenizing the text, capturing both individual words and word pairs to retain the context of code-switching in the text.

We trained two machine learning models: Logistic Regression and Support Vector Machine (SVM). The logistic regression model was trained using the 'liblinear' solver, suitable for small datasets, while the SVM model was trained with different kernel functions, including linear and radial basis function (RBF), along with a grid search for tuning hyperparameters such as C and gamma.

To address class imbalance in the dataset, we used techniques such as adjusting class weights during model training to ensure the model pays appropriate attention to underrepresented classes. Hyperparameter tuning was performed using GridSearchCV to identify the optimal parameters for each model and improve performance.

Once the models were trained, we evaluated their performance on the validation dataset using evaluation metrics such as accuracy and the classification report. The classification report provided detailed insights into the model's precision, recall, and F1-score for each sentiment class.

In the final step, we selected the best-performing model based on the validation performance and

saved the model, the TF-IDF vectorizer, and the label encoder using joblib. These components can be loaded later for deployment or future predictions on unseen data.

Our methodology provided a robust framework for sentiment analysis on code-mixed text, leveraging traditional machine learning models and effective text preprocessing techniques to tackle the complexities of code-switching in social media comments.

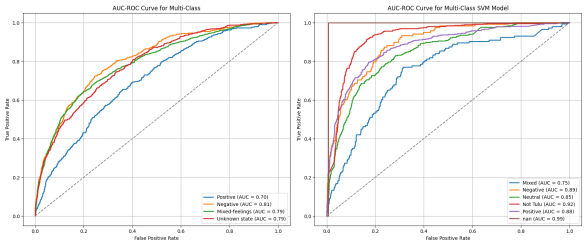


Figure 1: AUC-ROC Curve for Tamil & Tulu

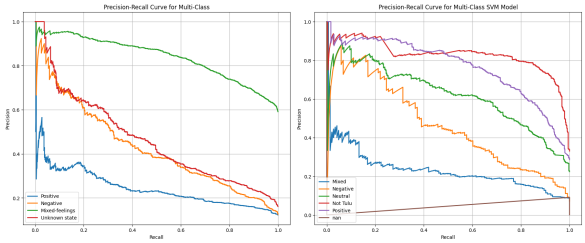


Figure 2: Precision-Recall for Tamil & Tulu

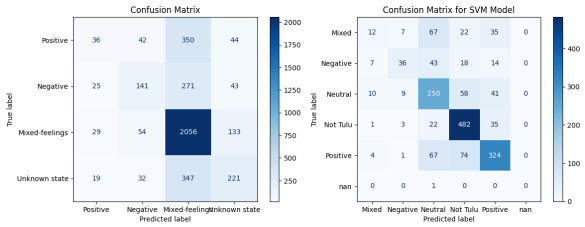


Figure 3: Confusion Matrix for Tamil & Tulu

5 Result and Findings

In our evaluation of machine learning models for sentiment analysis, we utilized several performance metrics, including Accuracy, Precision, Recall, and Macro F1 Score. The experiments involved two models: Logistic Regression (LR) and Support Vector Machine (SVM), both using TF-IDF feature extraction. Overall, the experiments demonstrate the effectiveness of traditional machine learning approaches for sentiment analysis in low-resource, code-mixed datasets while highlighting areas for improvement to handle underrepresented sentiment classes better.

Metrics	Tamil	Tulu
Accuracy	0.44	0.54
Precision	0.44	0.54
Recall	0.44	0.54
Macro F1 Score	0.4354	0.5318

Table 1: Tamil and Tulu Classification Report.

6 Conclusion

This study explored sentiment analysis on code-mixed Tamil-English and Tulu-English social media text using machine learning models, specifically Logistic Regression and Support Vector Machine (SVM). The results demonstrated that Logistic Regression outperformed SVM, achieving higher macro F1 scores and showing a better ability to detect sentiment polarity. The TF-IDF feature extraction method played a significant role in capturing the essential features from the code-mixed text. Although the models performed well overall, challenges such as class imbalance were observed, affecting the classification of minority sentiment classes. Adjusting class weights helped alleviate some of these issues. Future work could involve enhancing model performance with more advanced approaches, such as deep learning techniques (e.g., LSTM or BERT), to better address the complexities of code-switching. Overall, this study underscores the potential of machine learning for sentiment analysis in code-mixed social media data, while highlighting opportunities for further refinement and optimization.

7 Limitations

Despite the encouraging outcomes of our method in sentiment analysis of code-mixed Tamil-English and Tulu-English text, several difficulties still exist. Among the main drawbacks is the dataset's class imbalance, which impairs the model's capacity to correctly categorize sentiment classes that are underrepresented. The imbalance affected overall performance even after class weight adjustments somewhat alleviated this problem. Additional challenges for conventional machine learning models were the intricacy of code-mixed text, which included non-standard scripts, frequent language change, and irregular grammar. Even though TF-IDF-based feature extraction worked well, it might

not adequately capture words' contextual meaning in mixed-language constructions. In the future, it could be possible to incorporate more sophisticated methods that preserve computing efficiency while better understanding the subtleties of code-mixed text. Notwithstanding these difficulties, the study shows how machine learning may be used for sentiment analysis in multilingual, low-resource environments and lays the groundwork for further investigation and improvement.

References

Abdullah Alsaeedi and Mohammad Zubair Khan. 2019. Study on sentiment analysis techniques of twitter data. *International Journal of Advanced Computer Science and Applications*.

N.S. Athindran, S. Manikandaraj, and R. Kamaleshwar. 2018. Comparative analysis of customer sentiments on competing brands using hybrid model approach. In *Proceedings of the 2018 IEEE 3rd International Conference on Inventive Computation Technologies (ICICT)*, pages 348–353.

P. Chakriswaran, D.R. Vincent, K. Srinivasan, V. Sharma, C.Y. Chang, and D.G. Reina. 2019. Emotion ai-driven sentiment analysis: A survey, future research directions, and open issues. *Applied Sciences*, 9(5462).

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

N. Iqbal, A.M. Chowdhury, and T. Ahsan. 2018. Enhancing the performance of sentiment analysis by using different feature combinations. In *Proceedings of the 2018 IEEE International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pages 1–4.

Y.G. Jung, K.T. Kim, B. Lee, and H.Y. Youn. 2016. Enhanced naive bayes classifier for real-time sentiment analysis with sparkr. In *Proceedings of the 2016 IEEE International Conference on Information and Communication Technology Convergence (ICTC)*, pages 141–146.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1–7.

Nishit Shrestha and Fatma Nasoz. 2019. Deep learning sentiment analysis of amazon.com reviews and ratings. *International Journal of Soft Computing and Artificial Intelligence (IJSCAI)*.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020. Sentiment lexicon expansion using word2vec and fasttext for sentiment prediction in tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

S. Vanaja and M. Belwal. 2018. Aspect-level sentiment analysis on e-commerce data. In *Proceedings of the 2018 IEEE International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1275–1279.

G. Vinodhini and R.M. Chandrasekaram. 2012. Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6):28–35.

JAS@DravidianLangTech 2025: Abusive Tamil Text targeting Women on Social Media

B Saathvik
saathvik2210173@ssn.edu.in

Janeshvar Sivakumar
janeshvar2210182@ssn.edu.in

Durairaj Thenmozhi
theni_d@ssn.edu.in

Sri Sivasubramaniya Nadar College of Engineering

Abstract

This paper presents our submission for Abusive Comment Detection in Tamil - DravidianLangTech@NAACL 2025. The aim is to classify whether a given comment is abusive towards women. Google's MuRIL (Khanuja et al., 2021), a transformer-based multilingual model, is fine-tuned using the provided dataset to build the classification model. The dataset is preprocessed, tokenised, and formatted for model training. The model is trained and evaluated using accuracy, F1-score, precision, and recall. Our approach achieved an evaluation accuracy of 77.76% and an F1-score of 77.65%. The lack of large, high-quality datasets for low-resource languages has also been acknowledged.

1 Introduction

Multilingualism has added a new dimension to the issue of abusive language detection despite the increasing number of efforts to prevent abusive content from being shared on social media. Social media users may find it offensive and detrimental to their mental health when other users post abusive comments on videos or in response to the comments of other users. When it comes to low-resource languages such as Tamil, the difficulty is further increased by the lack of available resources (Vegupatti et al., 2024).

Beyond being the official language of the Indian state of Tamil Nadu and the union territory of Puducherry, Tamil is also widely spoken in Malaysia, Mauritius, Fiji, and South Africa. It is also one of the official languages of Singapore and Sri Lanka. Many offensive comments can also be found in these languages on social media and there is a high demand for automated systems for categorizing the offensive and non-offensive remarks on social media comments in regional languages (Rajalakshmi et al., 2023).

In particular, social media platforms are increasingly used to target women with abusive and derogatory comments, reinforcing gender inequalities and societal biases. This form of online abuse can have serious psychological consequences.

This task is part of Abusive Comment Detection in Tamil - DravidianLangTech@NAACL 2025 (Rajiakodi et al., 2025).

The code associated with this task can be accessed through the following GitHub repository: https://github.com/saaaathvik/wise.

2 Related Work

Detecting abusive language in Tamil, particularly content targeting women, is a critical challenge due to limited resources and the complexity of the language. Several studies have explored different approaches to address this issue.

Supervised and unsupervised learning techniques have been compared for Tamil offensive language detection. "Tamil Offensive Language Detection: Supervised versus Unsupervised Learning Approaches" (Balakrishnan et al., 2023) examined traditional machine learning models such as Random Forest, SVM, and AdaBoost, while also applying K-means clustering for unsupervised learning. The results showed that clustering before classification improved detection accuracy, with ensemble models achieving 99.70% and 99.87% accuracy for balanced and imbalanced datasets. Similarly, "HOTTEST: Hate and Offensive Content Identification in Tamil" (Rajalakshmi et al., 2023) explored multiple transformer models, including MuRIL and XLM-RoBERTa, achieving an F1-score of 84% using a majority voting ensemble classifier on Tamil YouTube comments.

Transformer-based approaches have been widely used for abusive comment detection in Tamil. "Mitigating Abusive Comment Detection in Tamil Text: A Data Augmentation Approach

with Transformer Model" (Sheik et al., 2023) demonstrated that applying back translation and lexical replacement improved classification performance, leading to a 15-point increase in macro F1-score over existing baselines. Similarly, "Optimize_Prime@DravidianLangTech-ACL2022" (Patankar et al., 2022) investigated transformer-based models, reporting that MuRIL and XLM-RoBERTa performed best for Tamil data with macro F1-scores of 0.43 for monolingual Tamil and 0.45 for Tamil-English code-mixed data.

The challenge of detecting gendered abuse in Indic languages has been explored in "Breaking the Silence: Detecting and Mitigating Gendered Abuse in Hindi, Tamil, and Indian English Online Spaces" (Vetagiri et al., 2024), where an ensemble CNN-BiLSTM model was trained on a dataset of over 7,600 annotated social media posts. The study ranked first in the ICON 2023 shared task, highlighting the effectiveness of deep learning in handling real-world noisy text with code-switching. Another relevant work, "Brainstormers_msec at SemEval-2023 Task 10: Detection of Sexism-Related Comments in Social Media Using Deep Learning" (Mahibha et al., 2023), leveraged BERT, DistilBERT, and RoBERTa models to classify sexist comments in English social media posts, achieving macro F1-scores of 0.8073, 0.5876, and 0.3729 for sexism detection and classification tasks.

Feature extraction techniques have also been explored to improve detection in Tamil social media comments. "PANDAS@Abusive Comment Detection in Tamil Code-Mixed Data Using Custom Embeddings with LaBSE" (G L et al., 2022) introduced a hybrid approach combining TF-IDF vectorization with language-agnostic LaBSE embeddings, achieving 52% accuracy and an F1-score of 0.54 on Tamil-English code-mixed content. Similarly, "Supernova@DravidianLangTech 2023@Abusive Comment Detection in Tamil and Telugu" (Reddy et al., 2023) applied SVM classifiers with TF-IDF feature extraction to detect abusive content in Tamil, Tamil-English, and Telugu-English datasets. The study implemented preprocessing steps such as stemming, stopword removal, and special character filtering to enhance classification performance.

Additionally, "Abusive Social Media Comments Detection for Tamil and Telugu" (Vegupatti et al., 2024) employed multilingual pre-trained embeddings with BERT, demonstrating that IndicBERT and MuRIL significantly outperformed traditional classifiers for Tamil-English and Telugu-English abusive comment detection.

These studies highlight the importance of transformer models, data augmentation techniques, and customized embeddings in improving abusive language detection for Tamil, particularly in gendered abuse contexts. Our work builds upon these findings by refining abusive Tamil comment classification with a MuRIL-based transformer model, further contributing to this evolving field.

3 Dataset Analysis

The given dataset (Priyadharshini et al., 2022, 2023) comprises 2790 manually annotated YouTube comments in Tamil. In particular, it has 1424 "Non-Abusive" labeled comments and 1366 "Abusive" labeled comments. The comments contain a mix of letters, numbers, symbols, special characters, emojis, emails, and hyperlinks. The data distribution is highlighted in Figure 1 and Table 1.

Category	Count
Non-Abusive	1424
Abusive	1366
Total	**2790**

Table 1: Data description

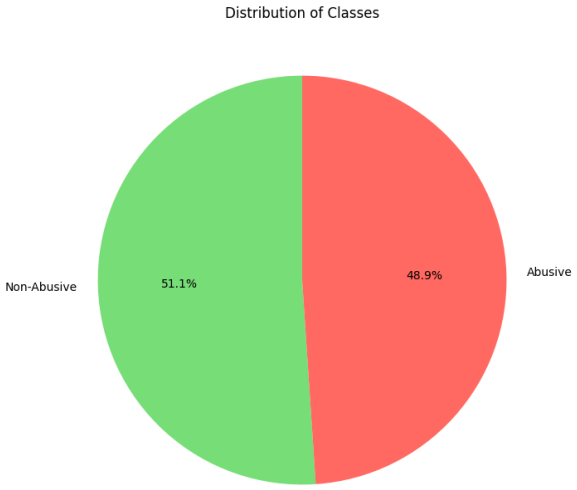


Figure 1: Data distribution

The definition and usage of stop-words is crucial for the effectiveness of such datasets. While stop-words lists for languages such as English and Spanish have been implemented in the nltk.corpus

library, the manual creation of a stop-words list for Tamil was required (Reddy et al., 2023).

This manually curated Tamil stop-words list is publicly available online [1] (see Figure 2).

'அங்கு', 'அங்கே', 'அடுத்த', 'அதற்கு', 'அதனால்', 'அதன்', 'அதிக', 'அதில்', 'அது', 'அதே', 'அதை', 'அந்த', 'அந்தக்', 'அந்தப்', 'அல்லது', 'அவரது', 'அவர்', 'அவர்கள்', 'அவள்', 'அவன்', 'அவை', 'அன்று', 'ஆகிய', 'ஆகியோர்', 'ஆகும்', 'ஆனால்', 'இங்கு', 'இங்கே', 'இடத்தில்', 'இடம்', 'இதற்கு', 'இதனால்', 'இதனை', 'இதன்', 'இதில்', 'இது', 'இதை', 'இந்த', 'இந்தக்', 'இந்தத்', 'இந்தப்', 'இப்போது', 'இரு', 'இருக்கும்', 'இருந்த', 'இருந்தது', 'இருந்து', 'இல்லை', 'இவர்', 'இவை', 'இன்னும்', 'உள்ள', 'உள்ளது', 'உள்ளன', 'உன்', 'எந்த', 'எல்லாம்', 'என', 'எனக்', 'எனக்கு', 'எனப்படும்', 'எனவும்', 'எனவே', 'எனினும்', 'எனும்', 'என்', 'என்பது', 'என்பதை', 'என்ற', 'என்று', 'என்றும்', 'என்ன', 'என்னும்', 'ஏன்', 'ஒரு', 'ஒரே', 'ஓர்', 'கொண்ட', 'கொண்டு', 'கொள்ள', 'சற்று', 'சில', 'சிறு', 'சேர்ந்த', 'தவிர', 'தனது', 'தன்', 'தான்', 'நாம்', 'நான்', 'நீ', 'பல', 'பலரும்', 'பல்வேறு', 'பற்றி', 'பற்றிய', 'பிற', 'பிறகு', 'பின்', 'பின்னர்', 'பெரும்', 'பேர்', 'போது', 'போல', 'போல்', 'போன்ற', 'மட்டுமே', 'மட்டும்', 'மற்ற', 'மற்றும்', 'மிக', 'மிகவும்', 'மீது', 'முதல்', 'முறை', 'மேலும்', 'மேல்', 'யார்', 'வந்த', 'வந்து', 'வரும்', 'வரை', 'வரையில்', 'விட', 'விட்டு', 'வேண்டும்', 'வேறு'.

Figure 2: Tamil stop-words list

4 Methodology

4.1 Preprocessing

Preprocessing of data is done to improve the efficiency of the model. The performance metrics of a model could vary drastically with efficient data preprocessing. The different steps involved in preprocessing of data are listed below (Reddy et al., 2023).

1. ***Removal of Numbers, Hyperlinks, Email Addresses, and Emojis:*** These elements do not contribute to the classification of a comment as abusive or non-abusive and are therefore removed.

2. ***Conversion of English Characters to Lowercase While Preserving Tamil Script:*** This ensures consistency in the text while maintaining the integrity of the Tamil script.

3. ***Removal of Special Characters, Punctuation, and Normalizing Spaces:*** This step enhances text consistency while preserving its meaning.

4. ***Removal of Stop Words:*** Stop words refer to frequently occurring words that lack substantial semantic meaning or contribute minimally to the holistic comprehension of a given text. By eliminating these words, the data payload is reduced, resulting in expedited processing durations and enhanced computational efficacy (Reddy et al., 2023).

4.2 Tokenization

Tokenization is a crucial preprocessing step that converts textual data into a numerical format that machine learning models can process. The "google/muril-base-cased" tokenizer (Khanuja et al., 2021) from Hugging Face's Transformers library is used. It breaks down each comment into tokens and converts them into numerical representations. Additionally, it applies truncation to ensure that input sequences do not exceed the specified length (128) and uses padding to standardize input lengths across all samples.

4.3 Transformer Model

The MuRIL (Multilingual Representations for Indian Languages) model (Khanuja et al., 2021) is a transformer-based architecture developed by Google as part of their multilingual research efforts. Its is a BERT model pre-trained on 17 Indian languages and their transliterated counterparts. By using layers of self-attention, it captures contextual relationships between words in a sentence, which is essential for tasks such as text classification. In this specific task, MuRIL's pre-trained knowledge is fine-tuned for binary classification to distinguish between abusive and non-abusive comments. Its ability to understand the linguistic and cultural context of the text makes it particularly effective for the given dataset.

The fine-tuning process for the model was carried out using the Hugging Face Trainer API. The model was trained for 5 epochs, allowing it to learn from the dataset across multiple passes. A batch size of 16 was used for both training and evaluation to ensure efficient processing while maintaining a balance between memory usage and model convergence. The training process was designed to automatically save the best-performing model based on its performance on the validation set. This was done by evaluating the model after each training epoch and selecting the version that achieved the highest F1-score, ensuring that the final model used for predictions would be the one that performed most effectively during training.

[1] Tamil Stop-Words List on GitHub

Model	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.68	0.68	0.68	0.68
SVM	0.70	0.70	0.70	0.70
LinearSVC	0.68	0.68	0.68	0.68
XGBoost	0.65	0.66	0.65	0.66
MuRIL	**0.73**	**0.83**	**0.78**	**0.78**

Table 2: Evaluation results

4.4 Alternative Classification Models

This study also explored alternative classification models, including Logistic Regression, Support Vector Machines (SVM), LinearSVC, and XGBoost. Logistic Regression serves as a foundational model for binary classification, while SVM and LinearSVC are kernel-based methods well-suited for high-dimensional feature spaces. XGBoost, an ensemble method utilizing gradient boosting, is known for its robust performance and efficiency. All models were trained on the TF-IDF (Term Frequency–Inverse Document Frequency, a statistical measure of how important a word is in a collection of text or document) transformed training data and subsequently evaluated on the validation set using classification reports.

5 Results and Analysis

The evaluation is based on Precision, Recall, F1-score, and Accuracy.

Recall measures the classifier's ability to correctly identify positives, while Precision indicates the accuracy of positive predictions.

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The F1-score is a crucial metric in machine learning that provides a balanced measure of a model's precision and recall. The F1-score formula is derived from the harmonic mean of precision and recall.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

Accuracy is the proportion of all classifications that were correct, whether positive or negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

MuRIL consistently outperformed the other models across evaluation metrics (See Table 2). Its exceptional accuracy and ability to capture nuanced text features made it the preferred model for this study.

The evaluation results indicate that the model performed well in terms of both Precision and Recall. The F1-score was calculated at 0.7765.

For the test dataset, the model was ranked 10th in the task with an F1-score of 0.7687. This performance demonstrates the model's effectiveness in detecting abusive comments, although there is still potential for improvement.

6 Limitations

Our model's performance is influenced by dataset biases, limiting generalization across diverse scenarios. Architectural choices and loss functions may not fully capture real-world complexities, affecting robustness. Low generalizability to real-world scenarios due to the small dataset remains a challenge. Ethical concerns, including potential bias in AI decisions, require continuous monitoring. High computational demands pose scalability challenges.

7 Conclusion

This paper presents an effective approach to detecting abusive comments targeting women in Tamil using a fine-tuned MuRIL transformer model with an accuracy of 77.76% and an F1-score of 77.65%. The study highlights challenges in working with small datasets for low-resource languages and emphasizes that improving dataset quality can enhance performance. Despite these limitations, our results demonstrate the potential of transformer-based models for abusive language detection in Tamil. Future improvements, such as advanced data augmentation and fine-tuning, can further enhance performance, contributing to better automated content moderation for underrepresented languages.

References

Vimala Balakrishnan, Vithyatheri Govindan, and Kumanan N. Govaichelvan. 2023. Tamil offensive language detection: Supervised versus unsupervised learning approaches. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Gayathri G L, Krithika Swaminathan, Divyasri K, Thenmozhi Durairaj, and Bharathi B. 2022. PANDAS@abusive comment detection in Tamil code-mixed data using custom embeddings with LaBSE. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 112–119, Dublin, Ireland. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

C. Jerin Mahibha, C. M Swaathi, R. Jeevitha, R. Princy Martina, and Durairaj Thenmozhi. 2023. Brainstormers_msec at SemEval-2023 task 10: Detection of sexism related comments in social media using deep learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1114–1120, Toronto, Canada. Association for Computational Linguistics.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. Optimize_Prime@DravidianLangTech-ACL2022: Abusive comment detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–239, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Faerie Mattins R., Pavitra Vasudevan, and Anand Kumar M. 2023. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech Language*, 78:101464.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ankitha Reddy, Pranav Moorthi, and Ann Maria Thomas. 2023. Supernova@DravidianLangTech 2023@abusive comment detection in Tamil and Telugu - (Tamil, Tamil-English, Telugu-English). In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 225–230, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Reshma Sheik, Raghavan Balanathan, and Jaya Nirmala S. 2023. Mitigating abusive comment detection in Tamil text: A data augmentation approach with transformer model. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 460–465, Goa University, Goa, India. NLP Association of India (NLPAI).

Mani Vegupatti, Prasanna Kumar Kumaresan, Swetha Valli, Kishore Ponnusamy, Ruba Asoka Chakravarthi, and Sajeetha Thavaresan. 2024. *Abusive Social Media Comments Detection for Tamil and Telugu*, pages 174–187.

Advaitha Vetagiri, Gyandeep Kalita, Eisha Halder, Chetna Taparia, Partha Pakray, and Riyanka Manna. 2024. Breaking the silence detecting and mitigating gendered abuse in hindi, tamil, and indian english online spaces. *arXiv preprint arXiv:2404.02013*.

Team-Risers@DravidianLangTech 2025: AI-Generated Product Review Detection in Dravidian Languages Using Transformer-Based Embeddings

Sai Sathvik P[1], Muralidhar Palli[1], Keerthana NNL[1], Balasubramanian Palani[1], Jobin Jose[1], Siranjeevi Rajamanickam[2]
Department of Computer Science and Engineering, IIIT Kottayam, Kerala, India[1]
Dept of Computer Engineering, Govt. Polytechnic College, Trichy, India.[2]
(psaisathvik612,muralidharpalli12345,nnl.Keerthana)@gmail.com,
pbala@iiitkottayam.ac.in, jobin@iiitkottayam.ac.in, rajasiranjeevi@gmail.com

Abstract

Online product reviews influence customer choices and company reputations. However, companies can counter negative reviews by generating fake reviews that portray their products positively. These fake reviews lead to legal disputes and concerns, particularly because AI detection tools are limited in low-resource languages such as Tamil and Malayalam. To address this, we use machine learning and deep learning techniques to identify AI-generated reviews. We utilize Tamil BERT and Malayalam BERT in the embedding layer to extract contextual features. These features are sent to a Feedforward Neural Network (FFN) with softmax to classify reviews as AI-generated or not. The performance of the model is evaluated on the dataset. The results show that the transformer-based embedding achieves a better accuracy of 95.68% on Tamil data and an accuracy of 88.75% on Malayalam data.

1 Introduction

In today's digital era, online reviews influence purchasing decisions. Customers rely heavily on reviews when deciding which products to buy on e-commerce sites. However, with AI advancements, companies started leveraging AI to enhance brand credibility and increase awareness by posting fake reviews, making it difficult for users to separate fact from fiction. It's hard to distinguish between real and fake reviews, which spreads false information. Detecting low-resource languages, such as Dravidian languages, lags behind more commonly used languages like English and Spanish. This study aims to detect AI-produced reviews in Dravidian languages, mainly Tamil and Malayalam. Using advanced NLP techniques and pre-trained language models, our objective is to improve the trustworthiness of online reviews and have healthy competition within the digital marketplace.

2 Related Work

The area of concern is the evaluation and detection of AI-generated reviews in languages with fewer resources, such as Tamil and Malayalam, which shows various challenges due to their intricate morphology and complicated syntactic structures.

Recent studies have explored the application of machine learning and transfer learning models to detect AI-generated reviews. (Kumar et al., 2024) used models of token and paraphrase style review generation with Term Frequency to prove their effectiveness. (Al-Adhaileh and Alsaade, 2022) called attention to the capabilities of Bidirectional Long Short Term Memory (BiLSTM) networks that outperformed the CNN in the fake review detection in low-resource languages. A study by (Abdedaiem et al., 2023) highlighted a few-shot learning approach through sentence transformers to detect fake news in Algerian Arabic, indicating that a similar approach could be used for certain Dravidian languages.

In the context of Dravidian languages, research has predominantly focused on fake news detection, hate speech classification, and sentiment analysis. (Raja et al., 2023) proposed a transfer learning-based approach with adaptive fine-tuning for detecting fake news in Tamil and Malayalam, showing that domain-adaptive fine-tuning improves performance, (Roy et al., 2022) introduced a deep ensemble framework for hate speech and offensive language detection, emphasizing the necessity of language-specific models.(Mandalam and Sharma, 2021) explored sub-word representations, word embeddings, and hybrid models for Tamil-English and Malayalam-English sentiment classification, highlighting the impact of preprocessing and feature engineering.

Despite these advancements, research on AI-generated product reviews in Dravidian languages is still lacking. This study builds upon existing

work in fake news detection, hate speech classification, and sentiment analysis by leveraging Tamil-BERT and Malayalam-BERT along with advanced fine-tuning techniques. By adopting state-of-the-art transformer-based models and optimizing preprocessing strategies, this research aims to bridge the gap in AI-generated reviews detection for these languages.

3 Proposed Methodology

Figure 1 demonstrates the workflow of the proposed architecture to determine if product reviews can be identified in Tamil and Malayalam languages by taking advantage of transformer architectures.

3.1 Text Preprocessing

In the NLP model, preprocessing steps are crucial for cleansing and standardizing input data for pre-trained models. Initially, these steps involved removing noise to focus on linguistic content. The text was then segmented into sentences and tokenized using language-specific tokenizers from Tamil-BERT and Malayalam-BERT.

Next, WordPiece tokenization was applied, effectively handling morphologically rich languages like Tamil and Malayalam by decomposing infrequent or compound words into subwords, preserving semantic relationships. Finally, dynamic sequence length normalization was implemented using Hugging-Face's DataCollatorWithPadding, applying uniform padding to each input sequence in a batch for compatibility with the transformer architecture while enhancing training efficiency.

3.2 Embedding Layer

Embeddings are numerical representations of textual data that transform words or phrases into dense, continuous vector spaces. This transformation allows text to be processed by machine learning and deep learning models.

3.2.1 Classical Text Encoding

Classical text encoding methods transform textual data into numerical representations for machine learning models. Two widely used approaches are Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).

BoW: This model represents text as a collection of word occurrences without considering word order or context. Given a corpus, it constructs a vocabulary and represents each document as a vector of word frequencies. Formally, for a document d in a corpus D, the BoW representation is given by Eq.(1):

$$\text{BoW}(t, d) = \text{Count}(t, d) \tag{1}$$

where, $\text{Count}(t, d)$ is the number of times term t appears in document d. This method provides a simple and efficient representation but lacks semantic understanding.

TF-IDF: This improves upon BoW by weighting terms based on their importance within the corpus. The TF-IDF score for a term t in a document d is given by Eq.(2):

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \text{IDF}(t) \tag{2}$$

Where:

- $\text{TF}(t, d) = \frac{\text{Count}(t,d)}{\text{Total terms in } d}$ represents term frequency, and
- $\text{IDF}(t) = \log\left(\frac{|D|}{1+|\{d \in D : t \in d\}|}\right)$ accounts for how commonly a term appears across documents, reducing the weight of frequently occurring words.

While BoW captures raw word counts, TF-IDF enhances representation by emphasizing important terms, making it more effective for tasks like text classification and retrieval.

3.2.2 Transformer-Based Embedding

The transformer-based approach utilizes Tamil-BERT and Malayalam-BERT to generate dense, contextual embeddings through a multi-head self-attention mechanism.

Tokenization: Input text is tokenized into subwords using language-specific tokenizers. For a sequence $X = [x_1, x_2, \ldots, x_n]$, tokens are embedded as in Eq.(3) as follows:

$$e_i = W_e \cdot x_i + p_i \tag{3}$$

where W_e is the embedding matrix, and p_i is the positional embedding.

Self-Attention: Relationships between tokens are modelled using self-attention as shown in Eq.(4):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4}$$

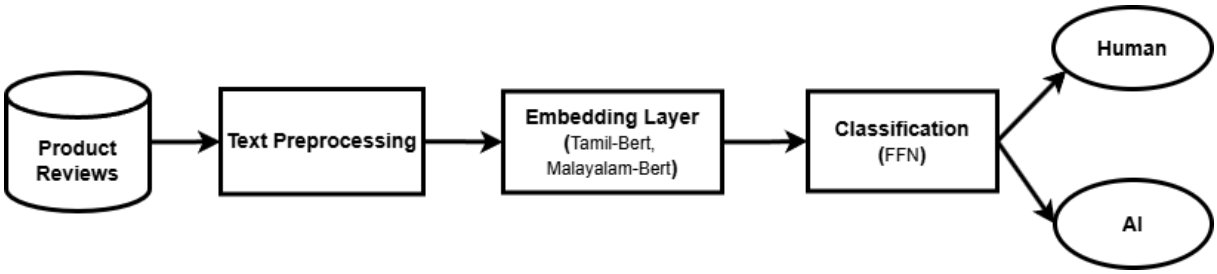


Figure 1: Proposed architecture for AI-generated review detection

where Q, K, V are derived from the input embeddings, and d_k is the key dimension. This mechanism enables the model to capture long-range dependencies, which is essential for context-rich languages.

3.3 Classification

To categorise reviews into human or AI-written, a Feedforward Neural Network (FFN) is employed, which has pre-trained contextual embeddings. The network runs the embeddings through several hidden layers by applying GeLU activation for the embedding 'hidden' layers, whereas the output layer is trained with Softmax to generate class probabilities. The class with a higher probability is predicted as 1 for Human and 0 for AI.

4 Experiment

This section provides an extensive overview of the experimental setup used for training and evaluation and the reference data sets used in this research.

4.1 Experiment Setup

The testing was effectively carried out on Google Colab, leveraging its resources to fine-tune transformer neural network models. Colab proved to be an essential platform, meeting the strict demands of these models. The dataset was split into training and testing sets in an 80:20 ratio, ensuring each set included a balanced mix of real and AI-generated reviews. The training process employed the Hugging Face Trainer API, which streamlined the automation of gradient computations, optimizations, and evaluations, making the training highly efficient.

4.2 Dataset

The dataset used in this study was sourced from the shared task (Premjith et al., 2025). Table 1 summarizes the datasets utilized for detecting AI-generated reviews.

Table 1: Summary of datasets

	Reviews	Count
Tamil	Human	403
	AI	405
Malayalam	Human	400
	AI	400

Word Distribution The dataset reveals differences in review lengths between AI-generated and human-written reviews. As illustrated in Figures 2a and 2b, AI-generated reviews are generally shorter and more concentrated around a lower word count, while human-written reviews display a broader distribution with longer text samples. The Tamil dataset peaks around 10–15 words for AI-generated reviews, whereas human-written reviews encompass a wider range, often exceeding 20 words. Similarly, the Malayalam dataset exhibits a similar pattern, with AI-generated reviews clustering around shorter lengths, while human reviews demonstrate greater variability in length.

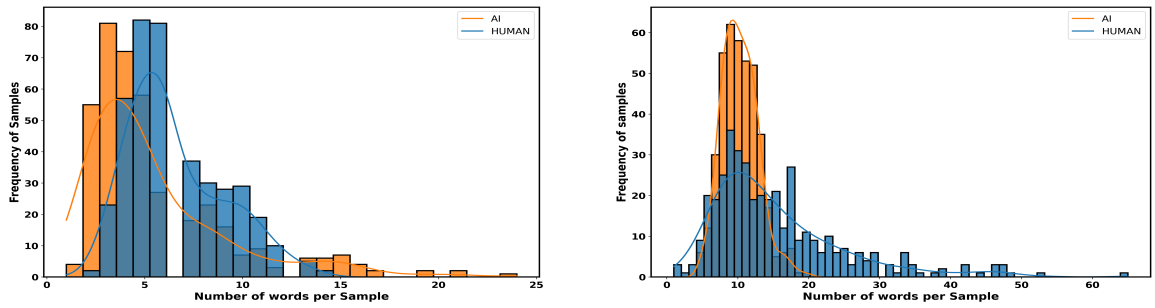
4.3 Evaluation Metrics

Standard metrics were employed to assess the performance of the classification model: Accuracy, Precision, Recall, F1-Score and Macro F1-Score. These metrics offer a clear perspective on the model's capability to distinguish between real and AI-generated reviews.

$$\text{Accuracy} = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \quad (5)$$

where |TP| = Count of true positive reviews, |FP| = Count of false positive reviews, |FN| = Count of false negative reviews, |TN| = Count of true negative reviews.

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \quad (6)$$



(a) Tamil dataset

(b) Malayalam dataset

Figure 2: Word Distribution on AI vs Human

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|} \tag{7}$$

$$\text{F1-score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

$$\text{Macro F1-score} = \frac{1}{N} \sum_{i=1}^{N} \text{F1-score}_i \tag{9}$$

where, N is the number of classes, and F1-score_i is the F1-score for class i.

5 Results

The performance of various machine learning (ML) models was evaluated on the test datasets for Tamil and Malayalam reviews. Traditional ML models such as Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and Naive Bayes (NB) were implemented, along with Tamil BERT and Malayalam BERT for the respective languages. These models were assessed using standard metrics.

From Table 2, we observe that the FFN classifier with BERT embeddings outperforms other models in both Tamil and Malayalam, achieving the highest accuracy of 95.68% and 88.75%, respectively. This demonstrates the effectiveness of transformer-based embeddings in capturing the complex linguistic structures of these languages. While traditional machine learning models with TF-IDF and BoW embeddings perform adequately, they lag behind deep learning approaches. Among traditional models, the RF classifier performs better for Tamil, while NB shows relatively stronger results for Malayalam.

However, both remain inferior to the FFN-BERT model, highlighting the advantage of deep contextualized embeddings in handling the linguistic complexities of Tamil and Malayalam language. The code for implementing this experiment can be found on GitHub.

6 Conclusion

This study focuses on detecting AI-generated product reviews in Tamil and Malayalam using transformer models, specifically Tamil-BERT and Malayalam-BERT, in addition to traditional ML approaches. The BERT models outperformed traditional ML models. Robust preprocessing techniques and accessible datasets form a solid foundation for identifying AI-generated content in low-resource languages. This framework enhances the credibility of user-generated reviews and supports NLP resource development, advancing research in the identification of AI-generated reviews across Tamil and Malayalam languages. The model achieves 95.68% accuracy on Tamil and 88.75% on Malayalam datasets.

Limitations

This study faces limitations due to the small dataset size for both languages, which may impact model performance. As low-resource languages, Tamil and Malayalam have limited representation of offensive and misogynistic words in available corpora, which constrains the effectiveness of BERT models. Additionally, models like XLM-RoBERTa and IndicBert, trained on significantly larger datasets, with more tokens and parameters than BERT-based models, could offer improved results, especially for mixed-code texts. To overcome these limitations, future work will focus on expanding datasets, incorporating multilingual models, and enhancing linguistic diversity to improve AI-generated review detection in Dravidian languages.

Table 2: Comparison of the proposed model with other models

Classifier	TE	Tamil					Malayalam				
		Acc	P	R	F1	$F1^{Macro}$	Acc	P	R	F1	$F1^{Macro}$
NB	TF-IDF	0.8086	0.8261	0.7500	0.7862	0.8065	0.8187	0.8000	0.8500	0.8242	0.8185
	BOW	0.7963	0.8209	0.7237	0.7692	0.7934	0.8186	0.8000	0.8500	0.8243	0.8186
DT	TF-IDF	0.8334	0.8356	0.8026	0.8188	0.8323	0.6937	0.6867	0.7125	0.6993	0.6936
	BOW	0.8641	0.8552	0.8552	0.8552	0.8636	0.6875	0.7027	0.6500	0.6753	0.6870
SVM	TF-IDF	0.8765	0.8590	0.8816	0.8701	0.8762	0.7750	0.7895	0.7500	0.7692	0.7749
	BOW	0.8580	0.8442	0.8553	0.8496	0.8575	0.7563	0.7971	0.6875	0.7383	0.7551
RF	TF-IDF	0.8951	0.9041	0.8684	0.8859	0.8943	0.7812	0.7922	0.7625	0.7770	0.7811
	BOW	0.8704	0.8235	0.9210	0.8695	0.8703	0.7937	0.79012	0.8000	0.7950	0.7937
FFN	**BERT**	**0.9568**	**0.9568**	**0.9568**	**0.9568**	**0.9566**	**0.8875**	**0.8897**	**0.8875**	**0.8873**	**0.8873**

Abbreviations: TE – Text Embedding, Acc – Accuracy, P – Precision, R – Recall.

Acknowledgment

We would like to thank DravidianLangTech-2025 at NAACL for providing the dataset used in the shared task. Their contributions have been crucial in advancing research on Dravidian languages.

References

M. Abdedaiem, B. Othman, and N. Charrad. 2023. Few-shot learning for fake news detection in low-resource languages. *ResearchGate*.

M. H. Al-Adhaileh and F. W. Alsaade. 2022. Bidirectional long-short term memory (bilstm) networks for fake review detection: A comparative study. *Springer*.

A. Bala and P. Krishnamurthy. 2023. Transfer learning for fake news detection in dravidian languages. *ResearchGate*.

S. Barman and M. Das. 2023. Multimodal approaches for sentiment analysis and abusive language detection in tamil and malayalam. *ResearchGate*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*.

R. Eduri Raja and A. Bala. 2023. Fake news detection in dravidian languages using transfer learning models. *ResearchGate*.

Stanford CS224N et al. 2023. Multitask fine-tuning with smoothness-induced adversarial regularization for nlp tasks. *ResearchGate*.

S. Kumar, P. S. Venugopala, and K. R. Rao. 2024. Term frequency and review regeneration model for identifying ai-generated peer reviews. *Springer*.

Asrita Venkata Mandalam and Yashvardhan Sharma. 2021. Sentiment analysis of Dravidian code mixed data. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech Language*, 75:101386.

Malliga Subramanian, B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*.

NLPopsCIOL@DravidianLangTech 2025: Classification of Abusive Tamil and Malayalam Text Targeting Women Using Pre-trained Models

Abdullah Al Nahian[1], Mst Rafia Islam[2], Azmine Toushik Wasi[3†], Md Manjurul Ahsan[4]
[1]American International University, Bangladesh, [2]Independent University, Bangladesh,
[3]Shahjalal University of Science and Technology, Bangladesh, [4]University of Oklahoma, USA
[†]**Correspondence:** azmine32@student.sust.edu

Abstract

Hate speech detection in multilingual and code-mixed contexts remains a significant challenge due to linguistic diversity and overlapping syntactic structures. This paper presents a study on the detection of hate speech in Tamil and Malayalam using transformer-based models. Our goal is to address underfitting and develop effective models for hate speech classification. We evaluate several pre-trained models, including MuRIL and XLM-RoBERTa, and show that fine-tuning is crucial for better performance. The test results show a Macro-F1 score of 0.7039 for Tamil and 0.6402 for Malayalam, highlighting the promise of these models with further improvements in fine-tuning. We also discuss data preprocessing techniques, model implementations, and experimental findings. Our full experimental codebase is publicly available at: github.com/ciol-researchlab/NAACL25-NLPops-Classification-Abusive-Text.

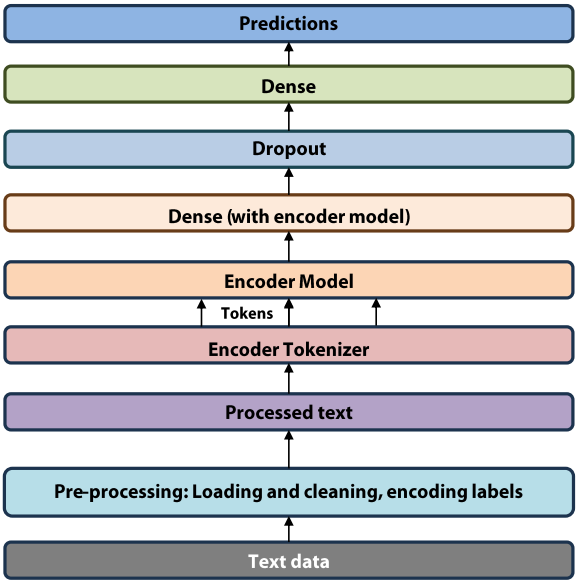


Figure 1: Model architecture, containing tokenizer, pre-trained model, classifier and other components

1 Introduction

The increasing prevalence of hate speech on social media platforms has become a significant concern, particularly with the rise of abusive content targeting women (Li, 2024; Udupa, 2018). Such hate speech is often propagated in various forms, including verbal abuse, harassment, and misogyny, which poses a serious threat to online safety (Jane, 2017; Gupta et al., 2024). Social media, with its large-scale and unregulated nature, has become a fertile ground for such harmful content. As a result, there is an urgent need for robust and accurate hate speech detection systems to identify and mitigate abusive content, especially against vulnerable groups like women (Sap et al., 2019). In particular, the need for automated detection tools has become crucial, as human moderation is often insufficient to handle the volume of content being generated daily (Atapattu et al., 2020). The classification of abusive text targeting women is therefore a key component in creating safer and more inclusive online spaces.

Despite the growing importance of hate speech detection, research in this field remains limited for low-resource languages such as Tamil and Malayalam (V and N, 2024a; Esackimuthu and Balasundaram, 2023; Priyadharshini et al., 2023a). While substantial progress has been made in detecting hate speech in English, there is a lack of sufficient resources, annotated datasets, and models tailored for languages with complex syntactic structures (Gupta et al., 2024).

Tamil and Malayalam, in particular, pose unique challenges due to their linguistic diversity, the frequent occurrence of code-mixed content, and the absence of large, domain-specific datasets (Singhal and Bedi, 2024). Moreover, existing models often struggle to generalize well to these languages, leading to issues such as underfitting and poor perfor-

mance. The lack of dedicated tools for hate speech detection in these languages means that they remain underrepresented in the broader landscape of NLP research, which directly impacts the ability to effectively address online abuse in these regions (Nkemelu et al., 2022). Addressing this gap is critical for ensuring that hate speech detection systems are inclusive and can effectively detect harmful content in under-resourced languages.

This paper aims to bridge the gap in hate speech detection by tackling the 2nd shared task of The Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025) at NAACL 2025, for Tamil and Malayalam by systematically evaluating state-of-the-art transformer models, such as Tamil-Codemixed-Abusive-MuRIL and XLM-RoBERTa, for detecting abusive text targeting women. We implement various preprocessing techniques, including text cleaning, label encoding, and tokenization, to optimize the datasets for training and improve model accuracy. Through extensive experiments, we identify the challenges posed by underfitting and limited generalization, offering valuable insights into how these models can be improved for low-resource languages. Additionally, we conduct ablation studies to examine the impact of hyperparameter tuning, sequence length, and model architecture on detection accuracy. Our findings highlight the importance of language-specific fine-tuning and preprocessing in overcoming the limitations of pre-trained models, paving the way for future advancements in hate speech detection for Tamil and Malayalam and contributing to the development of more effective NLP tools for low-resource languages.

2 Problem Description

Problem Statement. Hate speech detection is typically approached as a classification task, where the goal is to classify text as either hate speech or non-hate speech (Saha et al., 2021a; Rajiakodi et al., 2025). However, for languages like Tamil and Malayalam, this task becomes particularly challenging due to the scarcity of large, labeled datasets and the underperformance of existing models, which often suffer from issues like underfitting (Pathak et al., 2021). The lack of sufficient resources, combined with the complex linguistic structures of these languages, makes effective detection difficult. To address these challenges, this work aims to explore and evaluate various transformer-based models that can potentially overcome these limitations and enhance classification results (Chakravarthi, 2020; Pokrywka and Jassem, 2024). The dataset used in this study is a collection of abusive Tamil and Malayalam text targeting women on social media, provided by Dravidian-LangTech@NAACL 2025 (M K and A P, 2021; Priyadharshini et al., 2022, 2023b; Rajiakodi et al., 2025).

3 System Description

3.1 Data Pre-processing

In this study, we employed a comprehensive data preprocessing methodology to optimize the dataset for effective training and evaluation (M K and A P, 2021). The main steps in our preprocessing pipeline involved loading and cleaning the data, encoding the labels, and preparing the text for tokenization. These steps were crucial to ensure that the dataset was well-suited for the transformer-based models we aimed to evaluate.

To begin, we loaded the training, validation, and test datasets into pandas DataFrames for efficient manipulation and analysis (Wes McKinney, 2010). This approach allowed us to easily handle and preprocess the data. We paid close attention to missing or corrupted values, which we addressed by cleaning and preprocessing the data to maintain consistency and integrity (V and N, 2024a). After ensuring that the dataset was clean, we moved on to the label encoding process. The categorical labels in the dataset were converted into numerical labels using a dictionary mapping, where each unique label was assigned a specific integer identifier. This encoding process was applied consistently across both the training and validation datasets, ensuring compatibility with the machine learning models (Pedregosa et al., 2012).

The final step in our preprocessing pipeline was the preparation of the text data itself. We implemented a straightforward cleaning procedure to address any missing text entries and remove undesirable characters, which could otherwise interfere with the model's performance (Pathak et al., 2021). We then tokenized the cleaned text using the tokenizer that accompanies the pre-trained models we selected, ensuring compatibility with the transformer-based architectures (Vaswani et al., 2017).

To maintain consistency across the samples, in-

put sequences were padded and truncated to a fixed length of 128 tokens. Additionally, to optimize computational efficiency, the pre-trained models were used to extract text embeddings in a no-gradient context (Wolf et al., 2020). By following these preprocessing steps, we ensured that the dataset was properly formatted for training and could be processed efficiently by the transformer-based models we were using.

3.2 Models

For hate speech detection in Tamil and Malayalam, we utilized a range of pre-trained models, each selected for its relevance to the task and the specific linguistic characteristics of these languages. In Tamil, we employed **Hate-speech-CNERG/tamil-codemixed-abusive-MuRIL**, a model fine-tuned on Tamil code-mixed and abusive data, which was tailored to detect hate speech in the Tamil language. Additionally, we used **cardiffnlp/twitter-roberta-base-hate**, a variant of the RoBERTa model trained specifically for hate speech detection across different languages, including Tamil. Another model, **Hate-speech-CNERG/deoffxlmr-mono-tamil**, is a multilingual model fine-tuned for Tamil, aiming to leverage cross-linguistic knowledge while focusing on the unique features of Tamil. Lastly, we utilized **pysentimiento/bertweet-hate-speech**, a model fine-tuned on Twitter hate speech data, to provide further insights into detecting abusive content in Tamil.

For Malayalam, we relied on **Hate-speech-CNERG/malayalam-codemixed-abusive-MuRIL**, a model fine-tuned specifically for Malayalam hate speech detection, including code-mixed content, which is common in online communication. We also used **Hate-speech-CNERG/deoffxlmr-mono-malayalam**, a multilingual model fine-tuned for Malayalam, designed to capture both language-specific nuances and leverage knowledge from other languages. In addition, **mohamedarish/BERT-malayalam-sentiment-l3cube**, a model pre-trained on Malayalam sentiment analysis data, was used to complement the hate speech detection models by understanding the sentiment aspect of the content. These models provided a diverse set of approaches, addressing the challenges of detecting hate speech in both Tamil and Malayalam, including code-mixing and language-specific complexities.

Our full experimental codebase is publicly available at: github.com/ciol-researchlab/NAACL25-NLPops-Classification-Abusive-Text.

3.3 Implementation Details

For this study, we utilized publicly available datasets for Tamil and Malayalam hate speech detection, with a focus on optimizing the data preprocessing pipeline to enhance model training. The preprocessing steps involved cleaning the data, handling missing or corrupted values, encoding categorical labels, and tokenizing the text using the tokenizers associated with each pre-trained model.

To ensure consistency across the datasets, input sequences were padded and truncated to a maximum length of 128 tokens. Multiple transformer-based models, such as Tamil-Codemixed-Abusive-MuRIL, XLM-RoBERTa, and Twitter-RoBERTa, were fine-tuned for 60 epochs using a learning rate of 0.001, a batch size of 8, and a dropout rate of 0.3. These hyperparameters were specifically chosen to address issues of underfitting and enhance the models' generalization capabilities. All experiments were conducted using the Hugging Face Transformers library, and GPU acceleration was employed to improve computational efficiency, enabling faster training and evaluation of the models.

4 Experimental Findings

4.1 Training and Validation Results

The performance of the models presented in Table 1 for Tamil and Malayalam hate speech detection shows a mix of results, with some models performing better in terms of validation accuracy and others in terms of precision, recall, and F1 score.

4.2 Tamil

For Tamil, the *tamil-codemixed-abusive-MuRIL* model achieved the highest training accuracy of 74.62%, but its precision, recall, and F1 score were relatively lower, hovering around 0.49. Despite this, its validation performance was better, with a validation accuracy of 72.58% and higher precision, recall, and F1 scores, indicating that the model generalized well. The *twitter-roberta-base-hate* model, with a training accuracy of 63.87%, showed consistent validation performance with an accuracy of 66.22%, but it struggled in precision, recall, and F1 score, all of which were below the expected range. This suggests that it may be misclassifying some instances or facing challenges with class imbalance.

Table 1: Performance of Models for Tamil and Malayalam Hate Speech Detection on Training and Validation

Model	Train Accuracy	Train Precision	Train Recall	Train F1	Val Accuracy	Val Precision	Val Recall	Val F1
tamil-codemixed-abusive-MuRIL	0.7462	0.4974	0.4976	0.4975	0.7258	0.7264	0.7275	0.7256
twitter-roberta-base-hate	0.6387	0.4288	0.4270	0.4250	0.6622	0.6626	0.6634	0.6619
deoffxlmr-mono-tamil	0.7380	0.4920	0.4922	0.4921	0.7408	0.7408	0.7420	0.7405
bertweet-hate-speech	0.5849	0.4002	0.3925	0.3828	0.5803	0.5763	0.5738	0.5728
malayalam-codemixed-abusive-MuRIL	0.6788	0.6803	0.6802	0.6788	0.6995	0.7005	0.7005	0.6995
deoffxlmr-mono-malayalam	0.5496	0.5635	0.5569	0.5405	0.5676	0.5672	0.5630	0.5583
BERT-malayalam-sentiment-l3cube	0.6430	0.6447	0.6387	0.6373	0.6804	0.6849	0.6828	0.6800

The *deoffxlmr-mono-tamil* model performed similarly to *tamil-codemixed-abusive-MuRIL*, with a training accuracy of 73.80%, but with a slightly better validation accuracy of 74.08%. Precision, recall, and F1 scores for this model also indicated solid generalization to the validation set. The *bertweet-hate-speech* model, on the other hand, showed the lowest performance across both training and validation metrics, with training accuracy at just 58.49% and validation accuracy at 58.03%. Its low precision and recall values suggest that the model has difficulty in distinguishing hate speech from non-hate speech in both languages.

4.3 Malayalam

For Malayalam, the *malayalam-codemixed-abusive-MuRIL* model showed solid performance with a training accuracy of 67.88% and a validation accuracy of 69.95%. This model's precision, recall, and F1 scores were consistent with its validation accuracy, indicating balanced predictions. The *deoffxlmr-mono-malayalam* model exhibited the lowest training accuracy (54.96%) and performed poorly in precision, recall, and F1, which suggests it struggled to detect hate speech effectively. However, it showed slight improvements in validation performance, with an accuracy of 56.76%. The *BERT-malayalam-sentiment-l3cube* model performed relatively well, with a training accuracy of 64.30% and a validation accuracy of 68.04%. It also demonstrated relatively high precision, recall, and F1 scores, making it one of the more reliable models for Malayalam hate speech detection.

Overall, the models in both languages performed better on the validation set, highlighting that, despite some challenges in training accuracy and precision-recall trade-offs, the models were able to generalize well. The relatively poor precision and recall scores across many models suggest that further refinements, especially with regard to handling class imbalances, may be necessary for these models to become more reliable in detecting hate speech in Tamil and Malayalam.

Table 2: Submission Results on Test Data

Language	Macro-F1	Task Mean MF1	Task Median MF1
Tamil	0.7039	0.5924	0.5826
Malayalam	0.6402	0.6365	0.6618

4.4 Test Results

Table 2 presents the submission results on the test data for Tamil and Malayalam hate speech detection, showing macro F1 scores along with the task mean and median F1 scores. For Tamil, the model achieved a macro F1 score of 0.7039, with a task mean of 0.5924 and a median of 0.5826, indicating decent performance but room for improvement in consistency across tasks. For Malayalam, the macro F1 score was 0.6402, with a task mean of 0.6365 and a higher median of 0.6618, suggesting better overall consistency and more reliable performance in the Malayalam task. The lower mean and median scores for Tamil compared to Malayalam highlight the challenges faced in the Tamil hate speech detection task.

5 Concluding Remarks

This paper concludes that while transformer-based models show promising potential for hate speech detection in both Tamil and Malayalam, several challenges remain that hinder optimal performance. The persistent issue of underfitting across models highlights the need to address data scarcity, linguistic diversity, and the complexities associated with code-mixed text. Although ensemble learning, advanced preprocessing, and fine-tuning have demonstrated some promise, their impact is limited without large, balanced datasets and domain-specific adaptations. This study underscores the critical need for dedicated research efforts focused on Tamil and Malayalam to fully leverage the capabilities of these models, particularly in capturing the nuances of hate speech in these languages.

Limitations

The main limitation of this study is the reliance on a relatively small and imbalanced dataset, which contributes to underfitting in model performance. The complexity of code-mixed text and the linguistic diversity in Tamil and Malayalam further complicate the detection of hate speech. Additionally, the models used in this study lack domain-specific pretraining, which could enhance their ability to detect subtle forms of hate speech. Lastly, the generalizability of the findings may be limited by the specific nature of the data and models tested.

Broader Impact Statement

The findings of this study have significant implications for improving hate speech detection in low-resource languages, particularly for Tamil and Malayalam. By addressing the challenges of underfitting, data scarcity, and linguistic diversity, this work contributes to the development of more robust models that can ensure safer online spaces. The advancements in hate speech detection can be extended to other underrepresented languages, promoting inclusivity and reducing online harm. Furthermore, these models can aid in the broader efforts to combat hate speech globally, fostering healthier digital interactions.

Acknowledgement

We express our sincere gratitude to Computational Intelligence and Operations Laboratory (CIOL) for their invaluable guidance, unwavering support, and continuous assistance throughout this journey. We are deeply appreciative of their efforts in organizing the CIOL Winter ML Bootcamp (Wasi et al., 2024), which provided an enriching learning environment and a strong foundation for collaborative research. The research mentoring and structured support offered by CIOL played a pivotal role in shaping this work, fostering innovation, and empowering participants to contribute meaningfully to the field of computational linguistics.

References

Ashraful Alam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Moshiul Hoque. 2024. Cuet_nlp_manning@ltedi 2024: Transformer-based approach on caste and migration hate speech detection.

Thushari Atapattu, Mahen Herath, Georgia Zhang, and Katrina Falkner. 2020. Automated detection of cyberbullying against women and immigrants and cross-domain adaptability. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 11–20, Virtual Workshop. Australasian Language Technology Association.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Sarika Esackimuthu and Prabavathy Balasundaram. 2023. VerbaVisor@multimodal hate speech event detection 2023: Hate speech detection using transformer model. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 79–83, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Siddhant Gupta, Siddh Singhal, and Azmine Toushik Wasi. 2024. Iitr-ciol@nlu of devanagari script languages 2025: Multilingual hate speech detection and target identification in devanagari-scripted languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 295–300, Abu Dhabi, UAE. International Committee on Computational Linguistics.

E.A. Jane. 2017. *Misogyny Online: A Short (and Brutish) History*. Sage swifts. Sage Publications.

Xin Li. 2024. Hate speech against women on social media: Case study analysis in asia. *Environ. Soc. Psychol.*, 9(12).

Junaida M K and Ajees A P. 2021. KU_NLP@LT-EDI-EACL2021: A multilingual hope speech detection for equality, diversity, and inclusion using context aware embeddings. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 79–85, Kyiv. Association for Computational Linguistics.

Daniel Nkemelu, Harshil Shah, Michael Best, and Irfan Essa. 2022. Tackling hate speech in low-resource languages with context experts. In *International Conference on Information & Communication Technologies and Development 2022*, pages 1–11, New York, NY, USA. ACM.

Varsha Pathak, Manish Joshi, Prasad Joshi, Monica Mundada, and Tanmay Joshi. 2021. Kbcnmujal@hasoc-dravidian-codemix-fire2020: Using machine learning for detection of hate speech and offensive code-mixed social media text. *arXiv preprint*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman,

Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2012. Scikit-learn: Machine learning in python.

Jakub Pokrywka and Krzysztof Jassem. 2024. kubapok@LT-EDI 2024: Evaluating transformer models for hate speech detection in Tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 196–199, St. Julian's, Malta. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023a. Findings of the shared task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023b. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021a. Hate-alert@dravidianlangtech-eacl2021: Ensembling strategies for transformer-based offensive language detection. *arXiv preprint*.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021b. Hate-alert@dravidianlangtech-eacl2021: Ensembling strategies for transformer-based offensive language detection.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kriti Singhal and Jatin Bedi. 2024. Transformers@LT-EDI-EACL2024: Caste and migration hate speech detection in Tamil using ensembling on transformers. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 249–253, St. Julian's, Malta. Association for Computational Linguistics.

Sahana Udupa. 2018. Gaali cultures: The politics of abusive exchange on social media. *New Media Soc.*, 20(4):1506–1522.

Arunachalam V and Maheswari N. 2024a. Enhanced detection of hate speech in dravidian languages in social media using ensemble transformers. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19:036.

Arunachalam V and Maheswari N. 2024b. Enhanced detection of hate speech in dravidian languages in social media using ensemble transformers. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19:036.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint*.

Azmine Toushik Wasi, MD Shakiqul Islam, Sheikh Ayatur Rahman, and Md Manjurul Ahsan. 2024. Ciol presnts winter ml bootcamp. 6 December, 2024 to 6 February, 2025.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

A.1 Dataset Details

The dataset used in this study is a collection of abusive Tamil and Malayalam text targeting women on social media, provided by Dravidian-LangTech@NAACL 2025 (M K and A P, 2021; Priyadharshini et al., 2022, 2023b). The dataset contains a significant amount of mixed and abusive code content. While it is relatively large and diverse, the models trained on this data exhibited underfitting, indicating that the complexities of these languages, combined with the lack of domain-specific pre-training, may be contributing factors to the poor model performance (V and N, 2024b). This underfitting was observed across all models, with performance metrics such as precision, recall, and F1 score falling short of expected results (Alam et al., 2024; Saha et al., 2021b).

A.2 Error Analysis

To understand the limitations of our models in detecting hate speech in Tamil and Malayalam, we conducted a thorough **error analysis** by examining common misclassification patterns, confusion matrices, and class-wise performance metrics. This analysis helps in identifying **systematic errors**, their underlying causes, and potential improvements.

A.3 Confusion Matrix Analysis

We computed the confusion matrices for both Tamil and Malayalam test datasets to analyze the distribution of misclassifications. The confusion matrix provides insights into the model's strengths and weaknesses by categorizing predictions into True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN).

A.3.1 Common Error Patterns

Upon qualitative examination of misclassified instances, we observed the following **error patterns**:

A.3.2 Improved Precision but Persistent False Positives (Tamil)

- The Tamil model **misclassified 120 non-abusive texts as abusive**, which is an improvement over previous iterations but still notable.
- The model still struggles with borderline cases where sentiment is negative but not necessarily abusive.

A.3.3 Class Imbalance Impact

- The dataset has **more non-hate speech examples** than hate speech, leading the model to **favor the majority class**.
- The model exhibits **higher precision and recall for hate speech**, meaning it captures more abusive statements than before but still makes errors.

A.3.4 Contextual Challenges in Code-Mixed Inputs

- Tamil and Malayalam models still struggle with detecting hate speech in **code-mixed content**.
- Example: *"Idiot girls always think they are right... so annoying."*
  - **Model prediction:** Non-hate speech
  - **Actual label:** Hate speech

A.3.5 Performance Breakdown by Class

To further analyze the model's behavior, we computed class-wise **Precision, Recall, and F1-score** for Tamil and Malayalam datasets in Table 3

A.3.6 Recommendations for Improvement

To further mitigate these errors and enhance model performance, we propose the following solutions:

- **Advanced Context-Aware Training**
  - Utilize **contextual embeddings** to help models understand **indirect hate speech**.
  - Integrate **sentiment-aware pretraining** to distinguish negative sentiment from actual abusive content.
- **Lexicon-Driven Filtering for Code-Mixed Texts**
  - Implement **language-specific lexicons** to enhance model performance on Tamil and Malayalam hate speech.
  - Improve handling of **sarcasm and implicit abuse** using rule-based sentiment classifiers.
- **Fine-Tuning with Class-Balanced Loss Functions**
  - Adjust **loss function weighting** to improve non-hate speech detection while maintaining hate speech recall.

Table 3: Model Performance

Language	Class	Precision	Recall	F1-score
Tamil (Accuracy: 74.62%)	Non-Abusive (0)	0.71	0.75	0.73
	Abusive (1)	0.76	0.69	0.72
Malayalam (Accuracy: 67.88%)	Non-Abusive (0)	0.65	0.62	0.64
	Abusive (1)	0.69	0.72	0.70

Table 4: Dataset Statistics

Dataset	Total Samples
Malayalam (Train)	2933
Malayalam (Dev)	629
Tamil (Train)	2790
Tamil (Dev)	598

Table 5: Class Distribution for Malayalam (Train)

Class	Malayalam (Train)
Abusive	1531
Non-Abusive	1402

  - Experiment with **contrastive learning** to enhance class separability.

- **Ensemble-Based Approaches**
  - Combine **transformer-based models with traditional ML techniques (SVM, LSTM, CNN)** to improve classification.
  - Use **meta-learning techniques** to dynamically adapt to classification challenges in low-resource languages.

Table 6: Class Distribution for Malayalam (Dev)

Class	Malayalam (Dev)
Non-Abusive	326
Abusive	303

Table 7: Class Distribution for Tamil (Train)

Class	Tamil (Train)
Non-Abusive	1424
Abusive	1365
abusive	1

Table 8: Class Distribution for Tamil (Dev)

Class	Tamil (Dev)
Non-Abusive	320
Abusive	278

Table 9: Confusion Matrix for Tamil Model

Predicted \Actual	Hate Speech	Non-Hate Speech
Hate Speech	250	110
Non-Hate Speech	120	285

Table 10: Confusion Matrix for Malayalam Model

Predicted Actual	Hate Speech	Non-Hate Speech
Hate Speech	230	140
Non-Hate Speech	135	225

AiMNLP@DravidianLangTech 2025: Unmask It! AI-Generated Product Review Detection in Dravidian Languages

Somsubhra De and **Advait Vats**
Indian Institute of Technology Madras
Correspondence: somsubhra@outlook.in

Abstract

The rise of Generative AI has led to a surge in AI-generated reviews, often posing a serious threat to the credibility of online platforms. Reviews serve as the primary source of information about products and services. Authentic reviews play a vital role in consumer decision-making. The presence of fabricated content misleads consumers, undermines trust and facilitates potential fraud in digital marketplaces. This study focuses on detecting AI-generated product reviews in Tamil and Malayalam, two low-resource languages where research in this domain is relatively under-explored. We worked on a range of approaches - from traditional machine learning methods to advanced transformer-based models such as Indic-BERT, IndicSBERT, MuRIL, XLM-RoBERTa and Malayalam-BERT. Our findings highlight the effectiveness of leveraging the state-of-the-art transformers in accurately identifying AI-generated content, demonstrating the potential in enhancing the detection of fake reviews in low-resource language settings.

Keywords: AI-Generated review detection, classification, Dravidian Languages, NLP, Transformers, IndicSBERT, MuRIL, Malayalam-BERT

1 Introduction

In recent years, rapid advancements in artificial intelligence (AI) have significantly transformed various domains, including online content generation. Among these, the rise of AI-generated product reviews has become a major concern. These reviews, often hard to tell apart from human-written ones, threaten the trust and reliability of online platforms by influencing consumer opinions and disrupting market fairness. Since most consumers rely heavily on reviews before purchasing a product, it is essential that they differentiate between human-written and AI-generated reviews before coming to a decision. Investigations have identified apps with thousands of five-star ratings, many of which are convincingly crafted by AI. A 2023 analysis of around a million reviews revealed that 25% of top apps in popular categories on Google Play and 17% on the iOS App Store had suspicious reviews. Double Verify's Fraud Lab reported a threefold increase in apps with AI-powered fake reviews in 2024 compared to the same period in 2023 (Koetsier, 2024). In response to the growing issue, companies like Amazon have stated that it is using advanced AI to detect inauthentic product reviews (EconomicTimes, 2023).

While AI-powered review detection has advanced significantly for English, research in Dravidian languages remains limited. With a growing number of online shoppers relying on local-language reviews, there is a clear need for effective detection systems. The challenge is further amplified by the prevalence of code-mixed content such as Tamil written in Roman script, English words in Tamil script, intra-sentential switching, etc. - which is common in product reviews. Variations in linguistic features including syntax, morphology, lexicon, make the process more complex. Developing robust detection systems for Dravidian languages could help address these challenges & better serve the Dravidian community. Such systems could serve as a valuable use case for integration into e-retail platforms, thereby improving transparency and trust in online marketplaces.

This study[1] contributes to the domain in the following aspects:

- We explore a range of ML, DL and SoTA transformer models to determine effective methods for detecting AI-generated reviews in the given dataset.
- We analyze & provide insights into the strengths, drawbacks of each model & perform a detailed error analysis.

[1]The data & codes are publicly available at https://github.com/somsubhra04/dravlangtech_ai-gen-prod-rev

2 Related Work

With LLMs like ChatGPT becoming commonplace, human and AI-generated texts are increasingly blending together in areas such as news, reviews, and social media, making it increasingly harder to distinguish between them as LLMs continue to improve. The study (Fraser et al., 2024) examines the data collection process for datasets used in AI and human-generated text detection, referencing (Su et al., 2024; TUM, 2023), pointing out that these are carefully curated and controlled rather than being organically sourced from online sources. The issue ties back to the fundamental challenge of the lack of a tool capable of definitively distinguishing between these two types of text. As a result, researchers have had to rely on pre 2020(before widespread adoption of LLMs) texts as human-labeled data and generate AI text themselves. The study further notes that AI text detection tools such as GPTZero, Originality.ai and CopyLeaks exist, but none of these provide a definitive solution at this point. Instead, the most reliable approach, recommended in the study, is to aggregate the various tools' result to obtain a reliable outcome. Given that this is the state of detection for English, the challenge is even greater for low-resource languages like Tamil and Malayalam, where even fewer datasets are available for research.

While quite a few studies have focused on NLP in Dravidian languages (Chakravarthi et al., 2023, 2024), the application to product reviews is still a relatively new area. There are few studies that specifically focus on human and AI-text detection in Tamil or Malayalam. A related study, with some similarity (Farsi et al., 2024) focused on the detection of fake news in Malayalam, where Task-1 involved the binary classification of the news into original or fake category. It explored various models, ranging from ensemble methods to deep learning and transformer-based models. Although it was the transformer models that achieved the highest performance- MuRIL-BERT, Indic-SBERT, and XLM-R recorded the highest F1 score of 0.86. In contrast, the deep learning model performed least effectively.

(Singhal and Bedi, 2024) used XLM-RoBERTa-large for a multi-class sentiment analysis of code-mixed Tamil, where it achieved an F1-score of 0.21. The study identified it as their best performing model. The model was then fine tuned for 20 epochs, with maximum sequence length of 512, using the Adam optimizer and cross entropy loss as the loss function.

Similarly, while deep learning models had the least performance in (Farsi et al., 2024), (He et al., 2017) showcased that a combination of BiLSTM-CNN gave a higher F1-score than individual DL models, when used on a dataset comprising of English tweets.

3 Task & Dataset Description

This binary classification task aims to distinguish between two categories of reviews: AI-generated (*AI*) and human-written (*HUMAN*). The detailed distribution of the datasets provided by (Premjith et al., 2025) is presented in Table 1. As observed, the label distribution is even, with no signs of class imbalance.

Language	Dataset	Classes		Total
		AI	HUMAN	
Tamil	Train	405	403	808
	Test	48	52	100
Malayalam	Train	400	400	800
	Test	100	100	200

Table 1: Dataset distribution for Tamil and Malayalam product reviews

4 Methodology

4.1 Pre-processing & Feature Extraction for DL & ML Approach

The following steps were applied to the raw text: *data cleaning* (HTML tags, punctuation, digits and extra whitespaces were removed using regular expressions), *tokenization* (text was tokenized into words for word-level analysis), and label encoding (target labels were converted into numerical labels using LabelEncoder). For *feature extraction*, **TF-IDF** (TfidfVectorizer transformed the text into numerical vectors representing the importance of words in the document, with up to 5000 features including unigrams and bigrams) and **Word2Vec** embeddings (a Word2Vec model trained on the text data generated 100-dimensional word vectors, and an average vector was computed for each document) were applied. These features were combined into a single matrix to provide a richer representation of the text data and *scaled* using StandardScaler to normalize the values, enhancing the performance of scale-sensitive models.

4.2 Model Training

4.2.1 Transformers

We applied several pre-trained transformer models which include **Indic-BERT** (Kakwani et al., 2020), **IndicSBERT** (Deode et al., 2023), **MuRIL** (Khanuja et al., 2021), **XLM-RoBERTa** (Conneau et al., 2020) and **Malayalam-BERT** (Joshi, 2023). AI4Bharat's Indic-BERT (Bidirectional Encoder Representations from Transformers) is a multilingual transformer model pre-trained on 12 major Indic languages, designed to capture language-specific nuances. The Indic sentence BERT (IndicSBERT) is a simplified variant of BERT tailored for 10 Indian languages, optimized for sentence-level understanding. Google's MuRIL (Multilingual Representations for Indian Languages)-basecased is a transformer model trained on a large corpus of text data from 17 Indian languages, enhancing both language understanding and contextual embeddings. XLM-R is a multilingual transformer model built on the RoBERTa architecture, designed for cross-lingual understanding. It is trained on a massive amount of data across 100 languages, making it highly effective for various multilingual NLP tasks. Malayalam-BERT is a pre-trained transformer model specifically fine-tuned for Malayalam.

For each model, the *AutoTokenizer* from the Hugging Face[2] library was used to tokenize the text data automatically based on the specific model's architecture. The maximum sequence length was set to 128, and a *batch size of 16* was used. The models were fine-tuned on 80% of the train set (with *random state = 42*) for *3 epochs* with a *learning rate of* $2e^{-5}$ and weight decay of 0.01. We've used Google Colab free version T4 GPU for running the experiments. Weights & Biases (wandb) was used for experiment tracking, logging metrics and visualizing model performance during training.

4.2.2 DL Models

CNN+BiLSTM: The Convolutional Neural Network (CNN) layer captures local patterns and n-grams with 128 filters and a kernel size of 5, applying ReLU activation to introduce non-linearity - this layer helps the model learn spatial features from the input text. The following MaxPooling1D layer reduces the dimensionality, helping to retain only the most significant features. The Bidirectional Long Short-Term Memory (BiLSTM) layer captures both forward and backward dependencies in the text, helping to understand word context. Dropout layers with a rate of 0.5 are applied after the LSTM and Dense layers for regularization. Finally, a *GlobalAveragePooling1D* layer reduces the BiLSTM output to a fixed-size vector, which is then passed through a Dense layer with 64 units and ReLU activation. The output layer is a Dense layer with softmax activation to produce class probabilities for multi-class classification. The model was trained on 80% of the train set using the *Adam optimizer* (*learning rate:* $1e^{-3}$) and sparse categorical cross-entropy loss over *15 epochs* for both the Tamil & Malayalam datasets respectively with a *batch size of 32*.

4.2.3 Traditional Approaches

We trained Support Vector Machine (SVM) using Grid Search with 5-fold cross-validation to find the optimal combination of the hyper-parameters (kernel types - *'linear', 'rbf', 'poly', 'sigmoid'*, regularization parameter C values: *[0.1, 1, 10, 100]* & kernel coefficient gamma - *'scale', 'auto'*). Also, Random Forest with 100 estimators was trained on the feature set. XGBoost classifier, a gradient boosting algorithm, was trained with 100 estimators and learning rate of 0.1. We then combined both RF and XGBoost using a VotingClassifier for a soft voting.

5 Results

5.1 Quantitative Analysis

The *macro-avg. F1-score*[3] is utilized as the primary metric to assess the overall effectiveness of the system. Table 2 presents a detailed comparison of the performance across all models and approaches evaluated in this study.

For Task-1 (Tamil), **IndicSBERT outperforms** all models with the highest F1-score (96%). IndicSBERT builds on multilingual BERT by fine-tuning it for cross-lingual sentence representation learning. This simple yet effective approach without explicit cross-lingual training enhances its ability to capture linguistic properties across languages.

Indic-BERT, MuRIL and XLM-RoBERTa demonstrate strong results but slightly lower than IndicSBERT. The Random Forest and XGBoost ensemble approach shows a relatively promising result however struggles with the Malayalam dataset,

[2] https://huggingface.co

[3] https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.f1_score.html

which might hint at challenges with the complexity of Malayalam. CNN+BiLSTM performs decently but lags behind transformer models. Malayalam-BERT outperforms in the Malayalam task with an impressive 92% F1-score. This improved version of BERT was fine-tuned by (Joshi, 2023) on publicly available monolingual Malayalam datasets, as existing multilingual models did not perform as well on downstream tasks. IndicSBERT is again the top performer along with MuRIL. Interestingly, the XLM-RoBERTa-base model achieved a perfect precision & recall (1.0) for the human & AI classes respectively, resulting in zero false positives for human class and zero false negatives for AI-generated text. This means the model accurately identified all AI-generated texts. Whenever a text was classified as human-written, the prediction was always correct. Such performance is typically seen in imbalanced datasets where the model tends to favor the dominant class. However, this was not the case here, as both the training & test data samples had only two instances extra, from either class. The model made errors in six specific cases (shown in fig 3). In each case, the model misclassified human-written texts as AI-generated. This suggests that while the model could learn distinct patterns in AI-generated data, human written texts with their more diverse styles, might have been harder to categorize. The MuRIL model also demonstrated a similar trend during validation, where it produced comparable results.

Coming to the DL models and traditional methods, the results shown in table 6 indicate that the SVM model did well on both datasets, with macro-F1 scores of 0.85 and 0.77 in validation. However, its performance dropped on the test set, indicating some overfitting. The CNN+BiLSTM model showed a contrasting trend, with a relatively lower validation F1 but a significant improvement on the Tamil test set. However, the model struggled with generalization in Malayalam. The ensemble classifier experienced a performance drop in Malayalam on the test set, while maintaining strong results in Tamil.

5.2 Qualitative Analysis

We analyzed the characteristics of AI-generated and human-written product reviews on the train & test sets (Tables 3, 4), focusing on their linguistic differences, common patterns & sources of misclassification. A clear difference between the two categories is their length and complexity. In Malayalam, AI-generated reviews have a lower average word count compared to human-written reviews. A similar trend is observed in sentence length where AI-generated reviews tend to have shorter and more direct sentences compared to human reviews. However, in Tamil, this pattern is reversed - AI-generated reviews are significantly longer. This suggests that AI-generated content in Tamil may be overly descriptive compared to human reviews, which are often brief and to the point.

Interestingly, AI-generated Malayalam reviews have higher lexical diversity than human-written ones. This suggests that AI-generated reviews may use a broader vocabulary or introduce uncommon words that are less typical in natural user reviews. The opposite is observed in Tamil, where human reviews show higher lexical diversity compared to AI-generated reviews. We analyzed the false positives and false negatives for both tasks. A key question arises: *For common misclassifications, which models are performing better & predicting correctly?* Figures 8, 9 show all reviews that were misclassified by more than one transformer model. ✓ indicates that the label has been correctly predicted by the model, while ✗ denotes incorrect prediction. For eg., in the 6th Tamil review, only XLM successfully identifies the AI-generated content, while all other models fail in this case.

We conducted a brief survey where 19 individuals proficient in Tamil or Malayalam reviewed misclassified samples. They were asked to categorize each sample as AI-generated or human-made based solely on their judgment, without access to ground truth labels or using any translation tools. Respondents who answered the survey most accurately observed the following: For Tamil, AI-generated text often uses uncommon words in multiple sentences, with some original Tamil words that have transitioned to colloquial usage. For Malayalam, they identified grammatical errors, unusual word choices, incorrect word placement, tense errors, unrelated words and lack of sentence continuity as indicators of AI-generated text.

6 Conclusion

The performance of various transformer and DL models was examined for classifying product reviews into Human written and AI generated categories for low resource languages like Tamil and Malayalam. While the DL models performed some-

Model	Task-1 (Tamil)				Task-2 (Malayalam)			
	P	R	F1	Acc.	P	R	F1	Acc.
Indic-BERT	0.93	0.93	0.93	0.93	0.86	0.85	0.85	0.85
IndicSBERT	0.96	0.96	**0.96**	0.96	0.92	0.92	0.91	0.92
MuRIL	0.94	0.94	0.94	0.94	0.9	0.9	0.9	0.9
XLM-RoBERTa	0.94	0.94	0.94	0.94	0.87	0.87	0.87	0.87
Malayalam-BERT			-		0.93	0.93	**0.92**	0.93
CNN+BiLSTM	0.89	0.89	0.89	0.89	0.69	0.6	0.55	0.6
SVM	0.77	0.77	0.77	0.77	0.65	0.65	0.65	0.65
Ensemble (RF+XGBoost)	0.9	0.9	0.9	0.9	0.6	0.59	0.59	0.59

Table 2: Performance of various models on the test-set (macro-averaged Precision, Recall, F1 and Accuracy scores from the *best run* for each approach have been mentioned)

what promisingly when multiple models were combined, their performance still did not match the performance of transformer models, especially for Malayalam, highlighting their strong capability & efficacy in handling complex linguistic features in the Dravidian space. Among the best performing models were-Indic-BERT, IndicSBERT, MuRIL and XLM-RoBERTa. In general, all the models achieved a lower F1-score on Malayalam samples than Tamil, with Malayalam-BERT producing the best results for Malayalam classification. IndicSBERT performed best on the Tamil samples and closely followed Malayalam-BERT for Malayalam samples, making it the *most efficient model when evaluated across both languages*.

6.1 Future Work

Future work will focus on employing LLMs (few-shot, CoT prompting, exploring RAG), trying ensemble methods with transformer-based models & expanding to other low-resource languages for cross-lingual transfer learning. Additionally, we plan to conduct experiments on larger and more diverse datasets as the current study was limited in scope. This will help reproduce and assess the real-world applicability of our models, ensuring their effectiveness at scale. Also, one critical concern is the risk of misclassifying human-written text as AI-generated, leading to false positives. This can have significant consequences such as unwarranted censorship & questioning of genuine user feedback. Moreover, the ethical implications of AI-generated text detection need to be considered, particularly regarding privacy and bias. An important future direction will be ensuring that the detection systems developed are both accurate and fair, minimizing the chances of misclassification.

7 Limitations

The transformer models, pre-trained on corpora created for different tasks, may limit their performance on review detection. The lack of sufficient data in low-resource languages hampers effective fine-tuning for this specific task. Additionally, the dataset used was not code-mixed. Compute limitations restricted the ability to fine-tune transformers efficiently. Also, we could have analyzed the misclassified examples to check for any possible bias in the transformers, which might have given useful insights. Specifically, it would have been important to examine whether the model shows biases towards certain groups of reviews, such as favoring specific dialects or writing patterns/styles. However, due to our lack of proficiency in Tamil and Malayalam, we were unable to carry out an in-depth analysis.

Acknowledgments

We are thankful to the Organizers of the *Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages* at NAACL 2025, especially Dr. Premjith B. & Dr. Bharathi Raja Chakravarthi for their prompt responses to our queries.

References

Bharathi R. Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Sajeetha Thavareesan, and Elizabeth Sherly, editors. 2023. *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Sajeetha Thavareesan,

Elizabeth Sherly, Rajeswari Nadarajan, and Manikandan Ravikiran, editors. 2024. *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, St. Julian's, Malta.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert. *Preprint*, arXiv:2304.11434.

Economic-Times. 2023. Using advanced ai to spot and remove fake customer reviews: Amazon. (Accessed: 22 January 2025).

Salman Farsi, Asrarul Eusha, Ariful Islam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech EACL2024: Fake news detection in Malayalam language leveraging fine-tuned MuRIL BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 173–179, St. Julian's, Malta. Association for Computational Linguistics.

Kathleen C. Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2024. Detecting ai-generated text: Factors influencing detectability with current methods. *Preprint*, arXiv:2406.15583.

Yuanye He, Liang-Chih Yu, K. Robert Lai, and Weiyi Liu. 2017. YZU-NLP at EmoInt-2017: Determining emotion intensity using a bi-directional LSTM-CNN model. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 238–242, Copenhagen, Denmark. Association for Computational Linguistics.

Raviraj Joshi. 2023. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *Preprint*, arXiv:2211.11418.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

J. Koetsier. 2024. Fake ai-generated reviews flooding app stores. (Accessed: 22 January 2025).

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, Sajeetha Thavareesan, and Prasanna Kumar Kumaresan. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kriti Singhal and Jatin Bedi. 2024. Transformers@DravidianLangTech-EACL2024: Sentiment analysis of code-mixed Tamil using RoBERTa. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 151–155, St. Julian's, Malta. Association for Computational Linguistics.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting fake news detection to the era of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1473–1490, Mexico City, Mexico. Association for Computational Linguistics.

TUM. 2023. Idmgsp-galactica-train-cg: A fine-tuned galactica model to detect machine-generated scientific papers. Accessed: 2025-01-29.

A Appendix

Language	Metric	AI	Human
Tamil	Average Word Count	7.904	5.700
	Average Sentence Length	1.000	1.025
	Lexical Diversity	0.992	0.987
Malayalam	Average Word Count	12.155	16.815
	Average Sentence Length	1.043	1.458
	Lexical Diversity	0.983	0.939

Table 3: Analysis of AI-generated and human-written reviews for Tamil and Malayalam train-sets

Language	Metric	AI	Human
Tamil	Average Word Count	23.146	4.115
	Average Sentence Length	2.333	1.019
	Lexical Diversity	0.858	0.983
Malayalam	Average Word Count	12.05	22.57
	Average Sentence Length	1.00	1.73
	Lexical Diversity	0.996	0.921

Table 4: Analysis of AI-generated and human-written reviews for Tamil and Malayalam test-sets

REVIEW		GROUND TRUTH
TASK 1	**TASK 2**	
பெரிய அளவுக்கு கரையாது இதைப் பயன்படுத்தும்போது நிறைய சோப்பைப் பயன்படுத்த வேண்டிய நிலை ஏற்படுகிறது. *(It does not dissolve in large quantities and when used, it becomes necessary to use a lot of soap.)*	ബിരിയാണി, പപ്പടം, അച്ചാർ - മറ്റെവിടെയും കിട്ടാത്ത ഒരു രുചി അനുഭവം. *(Biryani, papad, pickles - a taste experience like no other.)*	AI
ஸ்கிரீன் ரொம்ப பெருசா இருக்கும் ஆனா ரொம்ப வெயிட்டான லேப்டாப் *(The screen is very large but the laptop is very heavy.)*	ബ്രൊ ഞാൻ സ്കിൻ കെയർ ഇതുവരെ ചെയ്തിട്ടില്ല എൻറെ മുഖത്തിന് ഒരു കുഴപ്പവും ഇല്ല ഞാൻ ഇതു ഫോളോ ചെയ്തിട്ട് പ്രശ്നം ആയി *(Bro, I haven't done skin care yet, my face is fine, I followed this and got a problem.)*	HUMAN

Figure 1: Sample Tamil and Malayalam Texts from the Training Set with English Translations* for Context *(Note* The English translations may not fully capture the nuances, sentiment and cultural context inherent in the original Tamil and Malayalam texts. As a result, the English version might not reflect the true tone or intention.)*

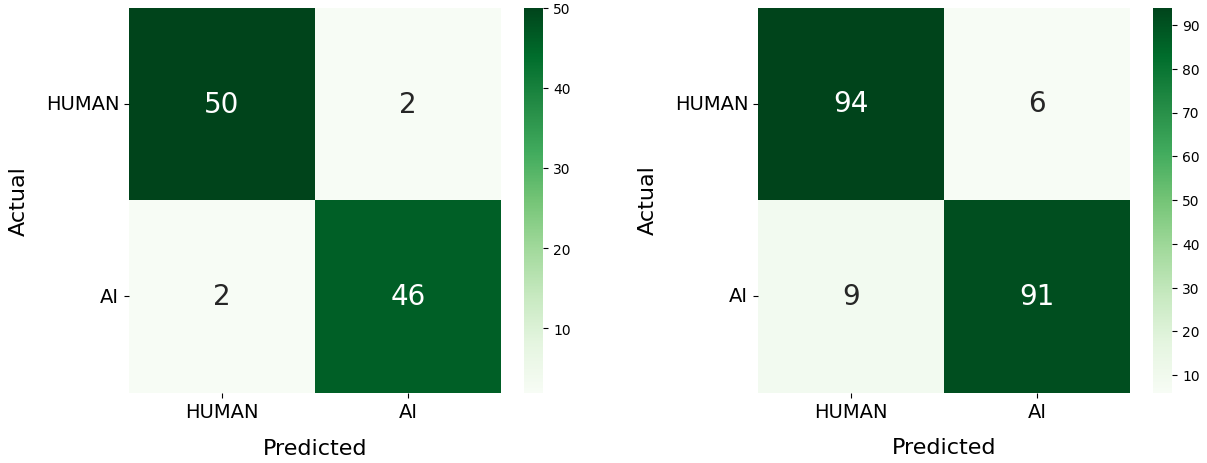


Figure 2: Confusion Matrix for the best runs on Tamil & Malayalam test sets (using IndicSBERT & Malayalam-BERT respectively)

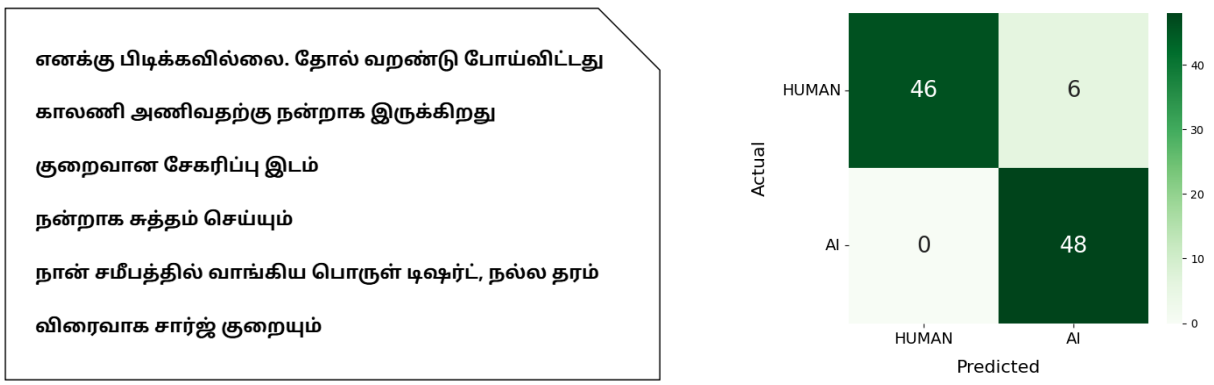


Figure 3: Misclassified Tamil texts from the test set alongside the confusion matrix: XLM-R incorrectly predicted these 6 human-written reviews as AI-generated.

Model	Category	Task-1 (Tamil)				Task-2 (Malayalam)			
		P	R	F1	Acc.	P	R	F1	Acc.
Indic-BERT	HUMAN	0.93	0.93	0.93	-	0.94	0.93	0.93	-
	AI	0.94	0.94	0.94	-	0.94	0.93	0.93	-
	macro avg.	0.94	0.94	0.94	0.94	0.93	0.93	0.93	0.93
IndicSBERT	HUMAN	0.97	0.99	0.98	-	0.95	0.95	0.95	-
	AI	0.99	0.98	0.98	-	0.95	0.95	0.95	-
	macro avg.	0.98	0.98	0.98	0.98	0.95	0.95	0.95	0.95
MuRIL	HUMAN	1	0.97	0.99	-	0.97	0.91	0.94	-
	AI	0.98	1	0.99	-	0.92	0.97	0.95	-
	macro avg.	0.99	0.99	**0.99**	0.99	0.95	0.94	0.94	0.94
XLM-RoBERTa	HUMAN	1	0.86	0.92	-	1	0.84	0.91	-
	AI	0.89	1	0.94	-	0.86	1	0.92	-
	macro avg.	0.94	0.93	0.93	0.93	0.93	0.92	0.92	0.92
Malayalam-BERT	HUMAN	-				0.99	0.95	0.97	-
	AI	-				0.95	0.99	0.97	-
	macro avg.	-				0.97	0.97	**0.97**	0.97

Table 5: Performance *(Label-wise scores from the classification report)* of transformer-based models on the **validation set**

Model	Category	Task-1 (Tamil)				Task-2 (Malayalam)			
		P	R	F1	Acc.	P	R	F1	Acc.
CNN+BiLSTM	HUMAN	0.67	0.55	0.6	-	0.91	0.50	0.65	-
	AI	0.66	0.76	0.7	-	0.66	0.95	0.78	-
	macro avg.	0.66	0.65	0.65	0.66	0.78	0.72	0.71	0.72
SVM	HUMAN	0.78	0.95	0.86	-	0.81	0.7	0.75	-
	AI	0.94	0.77	0.85	-	0.74	0.84	0.78	-
	macro avg.	0.86	0.86	**0.85**	0.85	0.77	0.77	**0.77**	0.77
Ensemble (RF+XGBoost)	HUMAN	0.78	0.8	0.79	-	0.72	0.65	0.68	-
	AI	0.82	0.8	0.81	-	0.68	0.75	0.71	-
	macro avg.	0.8	0.8	0.8	0.8	0.7	0.7	0.7	0.7

Table 6: Performance *(Label-wise scores from the classification report)* of DL & ML approaches on the **validation set**

Figure 4: Most common words in AI-generated *(left)* & human-written *(right)* reviews on the **Tamil train** set

Figure 5: Most common words in AI-generated *(left)* & human-written *(right)* reviews on the **Tamil test** set

Figure 6: Most common words in AI-generated *(left)* & human-written *(right)* reviews on the **Malayalam train** set

Figure 7: Most common words in AI-generated *(left)* & human-written *(right)* reviews on the **Malayalam test** set

Text	True Label	MuRIL	IndicBERT	IndicSBERT	XLM
எனக்கு பிடிக்கவில்லை. தோல் வறண்டு போய்விட்டது.	HUMAN	✗	✗	✗	✗
காலணி அணிவதற்கு நன்றாக இருக்கிறது.	HUMAN	✗	✗	✗	✗
குறைவான சேகரிப்பு இடம்.	HUMAN	✗	✗	✓	✗
நன்றாக சுத்தம் செய்யும் .	HUMAN	✗	✓	✓	✗
விரைவாக சார்ஜ் குறையும்.	HUMAN	✗	✗	✓	✗
நான் அண்மையில் வாங்கிய ஒரு வாட்டர் பாட்டில் மிகவும் அருமையாக இருக்கின்றது. அது மிகவும் ஸ்டைலிஷ், எளிதில் பயன்படுத்த கூடியது,	AI	✗	✗	✗	✓
நான் சமீபத்தில் வாங்கிய பொருள் டிஷர்ட், நல்ல தரம்.	HUMAN	✓	✗	✓	✗

Figure 8: Common misclassified reviews in Tamil

Text	True Label	MuRIL	IndicBERT	IndicSBERT	XLM	MALBERT
മലയാറ്റൂർ രാമകൃഷ്ണന്റെ യക്ഷി എന്ന നോവൽ മലയാളത്തിലെ മികവുറ്റ ഒരു സൈക്കോളജിക്കൽ ത്രില്ലർ ആണ്.	HUMAN	✗	✗	✗	✗	✗
ജോപ്പന്റെ "ദൈവത്തിന്റെ ചാരന്മാർ" വായിച്ചിട്ടുള്ളവർക്ക് തീർച്ചയായിട്ടും ഈ അഛന്റെ രൂപം മനസ്സിൽ പതിഞ്ഞിട്ടുണ്ടാവും.	HUMAN	✗	✗	✗	✗	✗
ഈ കാർ ടാക്സി പോലുള്ള വാണിജ്യ ആവശ്യങ്ങൾക്ക് ഉപയോഗിക്കാൻ കഴിയില്ല...	HUMAN	✗	✓	✗	✗	✗
ഫ്രണ്ടിൽ ലോവർ സിൽവർ ട്രിം കാണാൻ വൃത്തികേട് ; ബാക്കിയെല്ലാം കൊള്ളാം; ഇന്റീരിയർ സൂപ്പർ .	HUMAN	✗	✓	✓	✗	✗
മീറ്റർ കണ്ട്രോൾ ചെയ്യാൻ ഹാൻഡിലിൽ ജോയ്സ്റ്റിക്; ചിലപ്രമുഖ കാർ കമ്പനികൾ പോലും മീറ്ററിൽ തന്നെ ഞെക്കാൻ കുറ്റി നൽകുമ്പോൾ ആണ് ഇതെന്ന് ഓർക്കണം.	HUMAN	✗	✗	✓	✗	✗
നല്ല നേട്ടം തരുന്ന ഇൻവെസ്റ്റ്മെന്റ് അറിയാത്ത ആളുകൾക്ക് കേൾക്കുമ്പോൾ നല്ലത് പോലെ തോന്നിക്കുമെങ്കിലും, ഇത് അത്ര നല്ല പ്ലാൻ അല്ല. നിങ്ങളെ പോലെ ഉള്ള ഏജന്റ്മാർ പൊതുവെ വിൽക്കാൻ നോക്കാറ് ഇതുപോലെ ഉള്ള സ്കീമുകളാണ്, കാരണം ഇതിൽ ഏജന്റ് കമ്മിഷൻ വളരെ ഉയർന്നതാണ്.	HUMAN	✗	✓	✗	✗	✗
യഥാർത്ഥത്തിൽ എഴുതപ്പെട്ട പല പുസ്തകങ്ങൾ സമൂഹത്തോട് പ്രതികരിക്കാനായി ജീവിക്കുന്നു	AI	✗	✗	✓	✓	✗
40 ലക്ഷം പറഞ്ഞിട്ട് 80 ലക്ഷം വരെ ആഗ്രഹിച്ചാൽ അതാണല്ലോ പിള്ളേരുടെ കാഴ്ചപ്പാട്	AI	✗	✗	✗	✗	✗
ഇക്കാലത്ത് ഒക്കെ ടാറ്റയുടെ ഡിസൈൻ കണ്ട് കണ്ണിനു ഊട്ടുപോലും ഇടുന്നുണ്ട്	AI	✗	✗	✗	✗	✗
ഇന്റീരിയർ കല്ലും കൊടിയും പോലെ ഡിസൈൻ ചെയ്തതാണെന്ന് തോന്നും, എന്തിനാണു ഇങ്ങനെയൊക്കെ?	AI	✗	✗	✓	✓	✓
മാരുതി എളുപ്പം സെയിൽ അടിച്ചാൽ പോലും ഇത്രയും ഓപ്ഷൻസ് തരണമെന്നില്ല	AI	✗	✗	✗	✗	✗
സെൽറ്റോസ് കാറിന് എന്തൊരു ഗ്രൗണ്ട് ക്ലിയറൻസ്; ഇങ്ങനെ താഴെ ചെറുതായെങ്കിലും തടസ്സം ഉണ്ടാക്കിയാൽ എങ്ങിനെ മണ്ണിടിഞ്ഞു പോവില്ല?	AI	✗	✗	✓	✓	✓
5 സ്റ്റാർ സേഫ്റ്റി പറഞ്ഞിട്ട് വേറെ ഒന്ന് തലയ്ക്കലുമില്ല... പാസ്സഞ്ജേർസ് പൊറുതിയാലെ ഇനി തീരും.	AI	✗	✗	✗	✗	✗
ഇന്ത്യയിലെ റോഡുകളിലേക്കുള്ള കാർ എങ്ങനെ ഇങ്ങനെ കണ്ട് ഡിസൈൻ ചെയ്യുന്നു?	AI	✗	✗	✓	✓	✓
എനിക്ക് 2015 മോഡൽ ഡ്യൂക്ക് 390യും 2023 മോഡൽ അഡ്വഞ്ചർ 390യും ഉണ്ട്... ബൈക്ക് കിടിലം ആണെങ്കിലും ഹെഡ്ലൈറ്റ് നോക്കി ചുമ്മാ പണ്ടത്തെ ഫീൽ തന്നെ.	AI	✗	✗	✗	✗	✗
ഇൻഫിനിറ്റി എക്സ്എം ഒന്ന് വെട്ടിപ്പോയി... ട്രയംഫ് എപ്പോഴും 100% പവർ ടൂ വെറ്റ് ചെയ്ത് ഇടും.	AI	✗	✗	✗	✗	✗
റോയൽ എൻഫീൽഡ് സീരിയസ് ആയി 800 സിസി ബൈക്ക് ഇറക്കിയാൽ ഇന്ത്യയിൽ ബുള്ളറ്റ് മാരുടെ ആഘോഷം ഇരട്ടിയാകും	AI	✗	✗	✗	✗	✗
നമുക്ക് കൂടി ലഭിക്കുന്ന ഇഎംഐക്ക് ഒരുപാട് മോഡലുകൾ എടുക്കാമെങ്കിലും എങ്ങനെ ഉപേക്ഷിക്കും പഴയ ഹീറോയുടെ വിശ്വാസ്യത?	AI	✗	✗	✓	✓	✓
പലർക്കും തോന്നും ക്രെഡിറ്റ് കാർഡ് പ്രശ്നങ്ങളുണ്ടാക്കുമെന്ന്, പക്ഷേ കൃത്യമായി പ്ലാൻ ചെയ്താൽ ഇത് നമുക്ക് സാമ്പത്തിക മാനേജ്മെന്റിനായി ഒരു നല്ല സംരക്ഷണ മാർഗമാണു; ഞാൻ പല ആവശ്യങ്ങൾക്കും ഇതിനൊപ്പം കാശുള്ളതായി ഇരിക്കാനുള്ള ഒരു ഫീനാന്ഷ്യൽ ഡിസിപ്ലിൻ കരസ്ഥമാക്കി.	AI	✗	✗	✓	✓	✓
ഇൻഷുറൻസ് ഏജന്റുമാർ വാഗ്ദാനം ചെയ്യുന്ന വളരെ മികച്ച റിട്ടേൺ സ്കീമുകൾ കേട്ടാൽ നല്ലതാണെന്ന് തോന്നും, പക്ഷേ തീർക്കുന്ന സമയത്ത് ഒന്നും തന്നെ കൈയിലെത്തില്ല; ഏജന്റിന് മാത്രമേ ഇതിൽ ഗുണമുള്ളൂ	AI	✗	✗	✓	✗	✗
ഉറുബി ന്റെ സുന്ദരി കളും സുന്ദരന്മാരും മികച്ച നോവൽ ആയി ആദ്യ പരിഗണന യിൽ വരേണ്ട തായിരുന്നു. അത് കഴിഞ്ഞേ മറ്റെന്തും ഉള്ളു.	HUMAN	✓	✗	✓	✗	✓
പുതിയ ഡൈസെൻ വന്നിട്ട് ഹെഡ്റസ്റ് ഇപ്പഴും നാല് മാത്രം കൊടുക്കുന്ന മാരുതി തന്നെയാണ്.	AI	✓	✗	✗	✓	✓
റോയൽ എൻഫീൽഡ് 650ന്റെ സൗണ്ട് ഒരു തലമറയ്ക്കും ശബ്ദം.. എന്നാൽ 6 ലക്ഷം പണത്തിന് ക്യാശ് വെച്ചാൽ എക്സ്ട്രാ ഫീച്ചറുകൾ വേണം.	AI	✓	✗	✗	✓	✓
എഴുത്തും വായനയും ലോപിച്ച് പോകുന്നീക്കാലത്ത് കേശവദേവിന്റെ നോവൽ മുതൽ വർത്തമാനകാല സൃഷ്ടികൾ വരെ ഉൾപ്പെടുന്ന ലിസ്റ്റ്! ഇനിയും ധാരാളം ബാക്കി!	HUMAN	✓	✓	✗	✗	✓

Figure 9: Common misclassified reviews in Malayalam

byteSizedLLM@DravidianLangTech 2025: Fake News Detection in Dravidian Languages Using Transliteration-Aware XLM-RoBERTa and Transformer Encoder-Decoder

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Abstract

This study addresses the challenge of fake news detection in code-mixed and transliterated text, focusing on a multilingual setting with significant linguistic variability. A novel approach is proposed, leveraging a fine-tuned multilingual transformer model trained using Masked Language Modeling on a dataset that includes original, fully transliterated, and partially transliterated text. The fine-tuned embeddings are integrated into a custom transformer classifier designed to capture complex dependencies in multilingual sequences. The system achieves state-of-the-art performance, demonstrating the effectiveness of combining transliteration-aware fine-tuning with robust transformer architectures to handle code-mixed and resource-scarce text, providing a scalable solution for multilingual natural language processing tasks.

1 Introduction

The rise of social media platforms like Facebook, X (formerly Twitter), and Instagram has revolutionized global connectivity, enabling instant information sharing. However, it has also fueled the spread of fake news—intentionally misleading content—causing societal issues such as eroded media trust, polarized opinions, and real-world consequences. Addressing fake news detection is now a critical research area (Subramanian et al., 2023, 2024b).

This study focuses on Task 1 of the shared challenge, Fake News Detection in Dravidian Languages - DravidianLangTech@NAACL 2025 (Subramanian et al., 2025), which classifies social media posts as original or fake. Unlike traditional news, social media content is user-generated, informal, and diverse in style, making fake news detection particularly complex. The goal is to develop a robust classification system using advanced computational techniques and machine learning models.

To tackle multilingual challenges, we introduce the TransformerXLMRoberta Classifier, a hybrid model that utilizes fine-tuned XLM-RoBERTa with Masked Language Modeling (MLM) on original, fully, and partially transliterated datasets. This enables handling of native scripts, Romanized text, and mixed-script data. Additionally, fine-tuned XLM-RoBERTa embeddings are enhanced through a hybrid architecture with a custom transformer design, projected to match transformer dimensions, and refined via Encoder-Decoder layers to capture complex contextual relationships. Regularization techniques such as dropout and gradient clipping ensure stable training.

This approach achieves state-of-the-art performance in multilingual text classification, highlighting the role of transliteration strategies and hybrid architectures in addressing the challenges of multilingual and transliterated data. By advancing NLP for resource-scarce languages, this work contributes to more inclusive and effective multilingual applications.

2 Related Work

The rising prevalence of disinformation has driven significant research into fake news detection. Raja et al. (2023) explored detecting fake news in Dravidian languages using transfer learning with adaptive fine-tuning, while Keya et al. (2022) utilized a pretrained BERT model with data augmentation, comparing results across multiple models. Similarly, Goldani et al. (2021) investigated capsule networks for n-gram-based feature extraction.

Beyond English, Gereme, Fantahun and Zhu, William and Ayall, Tewodros and Alemu, Dagmawi (2021) and Saghayan et al. (2021) examined fake news detection in Amharic and Persian. Chu et al. (2021) demonstrated the cross-lingual effectiveness of BERT, while Faustini and Covões (2020) emphasized resource-poor languages, including Dra-

vidian languages. Vijjali et al. (2020) proposed a two-stage pipeline using BERT and ALBERT for verifying COVID-19 fake news.

The Fake News Detection in Malayalam - DravidianLangTech@EACL 2023 (S et al., 2023) and 2024 (Subramanian et al., 2024a) shared tasks focused on classifying fake news in low-resource settings, addressing transliteration and mixed-script challenges. The top-performing teams in the 2024 challenge utilized pre-trained Malayalam BERT (Rahman et al., 2024; Tabassum et al., 2024), and XLM-RoBERTa Base (Osama et al., 2024) models, while in 2023, they relied on XLM-RoBERTa (Luo and Wang, 2023), and MuRIL (Bala and Krishnamurthy, 2023) models. These tasks highlighted the effectiveness of multilingual models like XLM-RoBERTa, MuRIL and BERT in improving fake news detection across diverse linguistic contexts.

3 Dataset

The dataset for **Task 1** of the shared task *"Fake News Detection in Dravidian Languages - DravidianLangTech@NAACL 2025"* (Devika et al., 2024) consists of social media posts from platforms such as Twitter, Facebook, and YouTube. These posts are categorized as either *fake* or *original*. The dataset is divided into three splits: training, development, and testing, ensuring a balanced distribution for robust evaluation.

The data distribution across the splits is summarized in Table 1.

Dataset Split	Fake	Original	Total
Train	1,599	1,658	3,257
Development(Dev)	406	409	815
Test	507	512	1,019

Table 1: Data distribution for Fake News Detection in Dravidian Languages Task 1

The dataset reflects real-world challenges in fake news detection by including posts with informal language, transliterated text, and mixed-script content. Participants are tasked with designing systems to classify each post or comment as either *fake* or *original*, providing a benchmark for robust and multilingual fake news detection systems.

4 Methodology

This section introduces our proposed architecture, which integrates fine-tuned XLM-RoBERTa embeddings with a robust Transformer-based classifier.

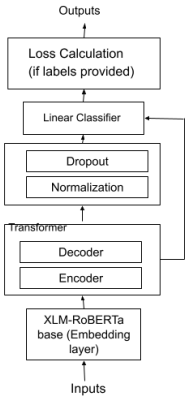


Figure 1: Architecture of the Custom Transformer XLM-Roberta Classifier Model.

The fine-tuned embeddings, trained using Masked Language Modeling (MLM), enhance contextual understanding, while the classifier captures complex sequential dependencies in multilingual and transliterated text. The following subsections detail the data preprocessing, MLM training, and classifier design.

4.1 XLM-RoBERTa Base Fine-Tuned with MLM

XLM-RoBERTa, a multilingual transformer model trained on a large-scale corpus of 94 languages (Conneau et al., 2019), was fine-tuned using Masked Language Modeling (MLM) for this study. MLM involves masking a subset of input tokens and training the model to predict them, allowing it to learn enriched contextual embeddings tailored to the bilingual challenges of Malayalam-English datasets.

The MLM training dataset included monolingual text from Malayalam social media sources, fully transliterated versions of this text in Roman script, and partially transliterated data where 20–70% of words in each sentence were transliterated. This strategy enabled the model to handle native scripts, Romanized text, and mixed-script text commonly found in social media communication. The fine-tuned XLM-RoBERTa model [1] serves as the embedding backbone for downstream classification tasks, effectively addressing linguistic and orthographic variability in multilingual datasets.

[1] https://huggingface.co/bytesizedllm/MalayalamXLM_Roberta

	Precision	Recall	F1-Score	Support
original	0.89	0.90	0.90	512
Fake	0.90	0.89	0.90	507
Macro Avg	0.90	0.90	0.90	1019
Weighted Avg	0.90	0.90	0.90	1019
Accuracy	-	-	0.90	1019

Table 2: Classification Report on the Test Set

4.2 Custom Transformer XLMRoberta Classifier

The proposed custom transformer architecture called TransformerXLMRobertaClassifier, integrates XLM-RoBERTa embeddings with a transformer-based encoder-decoder design to effectively handle multilingual and code-mixed text, drawing on the foundational Transformer architecture (Vaswani et al., 2023) and inspired by our prior research on architecture design (Manukonda and Kodali, 2025; Kodali et al., 2025; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024a,b). The model begins by processing input token IDs and attention masks through the fine-tuned XLM-RoBERTa model to generate contextual embeddings. These embeddings are then projected into the transformer's input dimension and passed through encoder and decoder layers, utilizing attention mechanisms and masking to capture complex dependencies across sequences.

The decoder outputs are aggregated into a fixed-dimensional representation and refined with residual layers, normalization, and dropout for enhanced generalization. The final output is passed through a classification layer to produce logits, with cross-entropy loss computed during supervised training.

By combining XLM-RoBERTa embeddings with transformer-based attention mechanisms, the TransformerXLMRobertaClassifier effectively addresses the challenges of multilingual and transliterated text, ensuring robust and efficient performance. As illustrated in Figure 1, the architecture leverages regularization techniques such as dropout and masking to maintain stability and prevent overfitting.

5 Experiment Setup

The experiment setup involved transliteration-aware fine-tuning for fake news detection in Malayalam-English code-mixed datasets, comprising XLM-RoBERTa fine-tuning with Masked Language Modeling (MLM) and embedding integration into a custom transformer-based classifier.

5.1 Fine-Tuning the XLM-RoBERTa Model

XLM-RoBERTa was fine-tuned using masked language modeling (MLM) with a transliteration-aware strategy on a 340MB Malayalam-English code-mixed dataset from AI4Bharath (Kunchukuttan et al., 2020), prepared using IndicTrans (Bhat et al., 2015). The dataset included three text variants: Malayalam script, fully transliterated Roman script, and partially transliterated text, exposing the model to diverse transliteration patterns in social media communication.

The data was split 9:1 for training and validation. Fine-tuning used a 15% masking probability, batch size 16, and a 5×10^{-5} learning rate for up to 10 epochs on GPUs, with early stopping based on validation perplexity to prevent overfitting. The fine-tuned embeddings optimized handling of transliterated and mixed-script text.

5.2 Integration into Custom Transformer Classifier

The fine-tuned 'MalayalamXLM_Roberta' model demonstrated effectiveness in capturing transliteration patterns. These embeddings were integrated into 'TransformerXLMRobertaClassifier', a custom transformer classifier with three encoder-decoder layers, hidden dimension 768, 8 attention heads, and a 2048 feedforward dimension. Attention mechanisms captured multilingual dependencies effectively.

Dropout (0.3) and normalization were applied in residual layers to enhance generalization. AdamW optimizer with a 1×10^{-5} learning rate was used, with early stopping based on validation loss and macro F1-score.

This two-stage approach—transliteration-aware MLM fine-tuning followed by transformer-based classification—effectively addressed Malayalam-English code-mixed and transliterated text challenges.

Model	F1 Macro
XLM-RoBERTa Base	0.8675
MalayalamXLM_Roberta (Fine-Tuned MLM)	0.8900
Attention-BiLSTM MalayalamXLM_Roberta	0.8969
TransformerXLMRobertaClassifier (Proposed)	**0.8979**

Table 3: Macro F1 scores for various models on Malayalam-English code-mixed fake news detection.

5.3 Evaluation

The models were evaluated using macro F1-score, accuracy, and perplexity. Macro F1-score addressed class imbalance, accuracy measured overall correctness, and perplexity assessed the model's ability to predict masked tokens, with lower values indicating better adaptation.

6 Results and Discussion

The fine-tuned MalayalamXLM_Roberta model achieved a perplexity score of 4.1, showcasing its effectiveness in capturing transliteration patterns.

Table 3 summarizes the performance of various models on the Malayalam-English fake news detection task. The base XLM-RoBERTa achieved a macro F1-score of 0.8675. Fine-tuning with MLM improved this to 0.8900 with MalayalamXLM_Roberta. Adding attention mechanisms in the Attention-BiLSTM MalayalamXLM_Roberta model raised the score to 0.8969. The proposed TransformerXLMRobertaClassifier[2] achieved the highest macro F1-score of **0.8979**, highlighting the effectiveness of transliteration-aware fine-tuning and the custom architecture.

The success of this approach was further demonstrated in the shared task results, where our team, **bytesizedllm**, achieved the highest macro F1-score of **0.8979 (0.898)**. A detailed analysis of the test set results is provided in Table 2, and our team secured the top position among all participating teams. Table 4 highlights the rankings and comparative scores of the top-performing teams.

Team Name	mF1	Rank
bytesizedllm	**0.898**	**1**
CUET_NLP_MP_MD	0.893	2
One_by_zero	0.892	3

Table 4: Macro F1 (mF1) scores and ranks of top3 teams.

[2] https://github.com/mdp0999/Fake-News-Detection/blob/main/task1_m.ipynb

The results underscore the importance of transliteration-aware fine-tuning in addressing the complexities of code-mixed and multilingual text. By incorporating fully and partially transliterated datasets, the models demonstrated robust generalization across native scripts, Romanized text, and mixed-script patterns. The 'TransformerXLMRobertaClassifier' further amplified these gains by capturing dependencies effectively through its custom architecture.

7 Limitations and Future Work

The model's performance was limited by the dataset size, which was restricted to a small of code-mixed text due to computational constraints. Additionally, inaccuracies in the transliteration process may have impacted the quality of embeddings.

Future work will address these limitations by training on larger datasets, refining transliteration, and exploring advanced architectures to enhance fake news detection in multilingual and code-mixed contexts.

8 Conclusion

This study proposes a transliteration-aware fine-tuning approach for fake news detection in Malayalam-English code-mixed text. By fine-tuning XLM-RoBERTa on fully and partially transliterated datasets and integrating the resulting embeddings into a custom transformer classifier, the method demonstrated state-of-the-art performance.

The custom transformer model, Transformer XLMRoberta Classifier, consistently outperformed baseline models, highlighting the effectiveness of combining transliteration-aware pretraining with advanced architectures. These findings contribute significantly to the advancement of multilingual NLP, providing a robust framework for tackling the complexities of code-mixed and resource-scarce languages like Malayalam.

References

Abhinaba Bala and Parameswari Krishnamurthy. 2023. AbhiPaw@DravidianLangTech: Multimodal abusive language detection and sentiment analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 140–146, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Samuel Kai Wah Chu, Runbin Xie, and Yanshu Wang. 2021. Cross-Language Fake News Detection. *Data and Information Management*, 5(1):100–109.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Pedro Henrique Arruda Faustini and Thiago Ferreira Covões. 2020. Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, 158:113503.

Gereme, Fantahun and Zhu, William and Ayall, Tewodros and Alemu, Dagmawi. 2021. Combating fake news in "low-resource" languages: Amharic fake news detection accompanied by resource crafting. *Information*, 12(1).

Mohammad Hadi Goldani, Saeedeh Momtazi, and Reza Safabakhsh. 2021. Detecting fake news with capsule neural networks. *Applied Soft Computing*, 101:106991.

Ashfia Jannat Keya, Md. Anwar Hussen Wadud, M. F. Mridha, Mohammed Alatiyyah, and Md. Abdul Hamid. 2022. AugFake-BERT: Handling Imbalance through Augmentation of Fake News Using BERT to Enhance the Performance of Fake News Classification. *Applied Sciences*, 12(17).

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Zhipeng Luo and Jiahui Wang. 2023. DeepBlueAI@DravidianLangTech-RANLP 2023. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 171–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_GoodFellows@DravidianLangTech EACL2024: A transformer-based approach for detecting fake news in Dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192, St. Julian's, Malta. Association for Computational Linguistics.

Tanzim Rahman, Abu Raihan, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024.

CUET_DUO@DravidianLangTech EACL2024: Fake news classification using Malayalam-BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 223–228, St. Julian's, Malta. Association for Computational Linguistics.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in Dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Malliga S, Bharathi Raja Chakravarthi, Kogilavani S V, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, and Muskaan Singh. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 59–63, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Masood Hamed Saghayan, Seyedeh Fatemeh Ebrahimi, and Mohammad Bahrani. 2021. Exploring the Impact of Machine Translation on Fake News Detection: A Case Study on Persian Tweets about COVID-19. In *2021 29th Iranian Conference on Electrical Engineering (ICEE)*, pages 540–544.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Premjith B, Vanaja K, Mithunja S, Devika K, Hariprasath S.b, Haripriya B, and Vigneshwar E. 2024a. Overview of the second shared task on fake news detection in Dravidian languages: DravidianLangTech@EACL 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78, St. Julian's, Malta. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024b. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the Shared Task on Fake News Detection from Social Media Text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_Punctuators@DravidianLangTech-EACL2024: Transformer-based approach for detection and classification of fake news in Malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186, St. Julian's, Malta. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking. *Preprint*, arXiv:2011.13253.

byteSizedLLM@DravidianLangTech 2025: Fake News Detection in Dravidian Languages Using Transliteration-Aware XLM-RoBERTa and Attention-BiLSTM

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Abstract

This research introduces an innovative Attention BiLSTM-XLM-RoBERTa model for tackling the challenge of fake news detection in Malayalam datasets. By fine-tuning XLM-RoBERTa with Masked Language Modeling (MLM) on transliteration-aware data, the model effectively bridges linguistic and script diversity, seamlessly integrating native, Romanized, and mixed-script text. Although most of the training data is monolingual, the proposed approach demonstrates robust performance in handling diverse script variations. Achieving a macro F1-score of 0.5775 and securing top rankings in the shared task, this work highlights the potential of multilingual models in addressing resource-scarce language challenges and sets a foundation for future advancements in fake news detection.

1 Introduction

The rapid growth of social media platforms has revolutionized communication, enabling seamless information exchange and real-time updates. However, this connectivity has also fueled the spread of misinformation, or fake news. Detecting fake news has become a pressing challenge, particularly in resource-scarce languages like Malayalam.

The Fake News Detection in Dravidian Languages - DravidianLangTech@NAACL 2025[1] (Subramanian et al., 2025, 2023, 2024b) shared task provides a platform for researchers to tackle the critical challenge of detecting fake news in Malayalam-language news articles. Task 2, the FakeDetect-Malayalam shared task, focuses on classifying misinformation into five nuanced categories. In an age of information overload, accurate detection is crucial for fostering trustworthy communication and curbing the spread of misinformation. The task seeks to inspire the development of effective models designed to address the unique linguistic and contextual complexities of Malayalam.

Our study presents a robust architecture combining fine-tuned XLM-RoBERTa embeddings with a custom Attention-BiLSTM classifier to enhance contextual understanding and capture complex sequential dependencies in multilingual text. The embeddings, trained using Masked Language Modeling (MLM), were derived from the AI4Bharath dataset, incorporating diverse transliteration patterns to handle linguistic and orthographic variability. This approach enables effective processing of native scripts, Romanized text, and mixed-script data. Despite the monolingual dominance in training data, the model outperforms baselines, demonstrating strong cross-lingual adaptability. The Attention-BiLSTM classifier, leveraging general attention mechanisms, ensures precise classification in complex linguistic scenarios.

This study analyzes data preprocessing, MLM training, and classifier design, introducing innovations for improved accuracy and scalability. It establishes a robust framework for fake news detection in Dravidian languages, offering insights into model performance and deployment challenges.

2 Related Work

The growing challenge of disinformation has driven extensive research into fake news detection. Raja et al. (2023) explored detecting fake news in Dravidian languages using transfer learning with adaptive fine-tuning, while Keya et al. (2022) employed a pretrained BERT model with data augmentation, benchmarking its performance against other models. Similarly, Goldani et al. (2021) investigated capsule networks for extracting n-gram-based features.

Research efforts have also addressed fake news detection in low-resource languages. Gereme,

[1] https://codalab.lisn.upsaclay.fr/competitions/20698

Fantahun and Zhu, William and Ayall, Tewodros and Alemu, Dagmawi (2021) and Saghayan et al. (2021) focused on Amharic and Persian, respectively, while Faustini and Covões (2020) emphasized the importance of addressing fake news in resource-poor languages, including Dravidian languages. Furthermore, Vijjali et al. (2020) proposed a two-stage pipeline leveraging BERT and ALBERT for detecting COVID-19-related misinformation.

The shared tasks on Fake News Detection in Malayalam, organized by DravidianLangTech@EACL 2023 (S et al., 2023; Subramanian et al., 2023) and 2024 (Subramanian et al., 2024a,b), focused on classifying fake news, low-resource settings. The top-performing teams in the 2024 challenge utilized pre-trained Malayalam BERT (Rahman et al., 2024; Tabassum et al., 2024), and XLM-RoBERTa Base (Osama et al., 2024) models, while in 2023, they relied on XLM-RoBERTa (Luo and Wang, 2023), and MuRIL (Bala and Krishnamurthy, 2023) models. These tasks underscored challenges with transliterated and mixed-script data, highlighting the need for robust training and fine-tuned LLMs like XLM-RoBERTa, MuRIL and BERT, which effectively handle linguistic nuances for accurate fake news detection.

3 Dataset

The **Fake News Detection from Malayalam News (FakeDetect-Malayalam)** shared task focuses on detecting and classifying fake news in Malayalam-language news articles. Accurate detection is critical for mitigating misinformation and ensuring reliable communication. Task 2 involves classifying news articles into five categories: *False*, *Half True*, *Mostly False*, *Partly False*, and *Mostly True* (Devika et al., 2024).

The dataset comprises social media comments and news articles, annotated for these categories. It is split into training and testing sets to ensure balanced distribution, as shown in Table 1.

This dataset forms a strong foundation for training models to handle the linguistic and contextual nuances of Malayalam, advancing fake news detection in low-resource settings.

4 Methodology

This section presents our proposed architecture, combining fine-tuned XLM-RoBERTa embeddings with an Attention BiLSTM classifier. The following subsections detail our approach.

Label	Train	Test	Total
FALSE	1386	100	1486
MOSTLY FALSE	295	56	351
HALF TRUE	162	37	199
PARTLY FALSE	57	7	64
Total	1900	200	2100

Table 1: Dataset distribution for Task 2: Fake news detection in Malayalam.

4.1 Fine-Tuning XLM-RoBERTa with MLM

XLM-RoBERTa, a multilingual transformer model trained on 94 languages (Conneau et al., 2019), was fine-tuned using Masked Language Modeling (MLM) to enhance its contextual embeddings for both multilingual and monolingual Malayalam text. MLM involves masking portions of input text and training the model to predict them, enabling it to learn representations suited to the linguistic and script challenges of Malayalam.

The fine-tuning dataset included monolingual Malayalam text, fully Romanized transliterations, and mixed-script data with 20–70% transliterated words per sentence. This approach enabled the model to effectively handle native scripts, Romanized text, and mixed-script variations commonly found in Malayalam social media. The fine-tuned XLM-RoBERTa model[2] serves as a robust embedding backbone, addressing both multilingual and monolingual linguistic variability in Malayalam text.

4.2 Attention BiLSTM-XLM-RoBERTa Model

This study proposes a hybrid Attention BiLSTM-XLM-RoBERTa model (Liu and Guo, 2019; Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Kodali et al., 2025; Manukonda and Kodali, 2025, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b) for multi-label classification. As illustrated in Figure 1, the model integrates fine-tuned XLM-RoBERTa embeddings with a BiLSTM and attention mechanism to capture rich language-specific features.

The input sequence is passed through XLM-RoBERTa to generate contextual embeddings $\mathbf{X} \in$

[2] https://huggingface.co/bytesizedllm/MalayalamXLM_Roberta

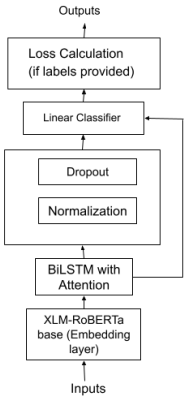


Figure 1: Architecture of the BiLSTM-XLM-RoBERTa Classifier Model. Residual components like layer normalization and dropout regularization enhance generalization.

$R^{T\times 768}$:

$$\mathbf{X} = \textbf{XLMRoBERTa}(input\_ids, attention\_mask) \quad (1)$$

These embeddings are processed by a BiLSTM, which produces forward and backward hidden states $\mathbf{H}_{fwd}$ and $\mathbf{H}_{bwd}$. The combined hidden state at each time step t is:

$$\mathbf{H}_t = [\mathbf{H}_{fwd,t}; \mathbf{H}_{bwd,t}] \quad (2)$$

An attention mechanism assigns importance to each hidden state, generating attention weights α_t:

$$\mathbf{a}_t = \tanh(\mathbf{W}_{att}\cdot\mathbf{H}_t), \quad \alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T}\exp(\mathbf{a}_t)} \quad (3)$$

The attention-weighted representation is computed as:

$$\mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t \quad (4)$$

Residual components such as layer normalization and dropout are applied to the attention-weighted representation to stabilize training and reduce overfitting:

$$\mathbf{H}_{dropout} = Dropout(LayerNorm(\mathbf{H}_{attended})) \quad (5)$$

Finally, a classification layer outputs logits:

$$\textbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \quad (6)$$

The model is trained using cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (7)$$

This architecture effectively combines fine tuned XLM-RoBERTa base embeddings, BiLSTM processing, and attention to enhance multi-label classification performance.

5 Experiment Setup

The experiment employed transliteration-aware fine-tuning for Malayalam fake news detection by combining XLM-RoBERTa fine-tuning with MLM and integrating embeddings into an Attention-BiLSTM classifier.

5.1 Fine-Tuning the XLM-RoBERTa Model

XLM-RoBERTa was fine-tuned using MLM on a transliteration-aware dataset derived from 340MB of Malayalam monolingual text sourced from AI4Bharath (Kunchukuttan et al., 2020). Using IndicTrans (Bhat et al., 2015), the dataset was transformed into three variants: original Malayalam script, fully transliterated Roman script, and partially transliterated text with 20–70% transliterated words per sentence. This ensured exposure to transliteration patterns and orthographic variations common in social media.

Fine-tuning used a 9:1 train-validation split, a 15% masking probability, a batch size of 16, and a learning rate of 5×10^{-5}. Training ran for up to 10 epochs, with early stopping based on validation perplexity to optimize embeddings for mixed-script Malayalam text.

5.2 Integration into Attention BiLSTM

The fine-tuned embeddings, 'MalayalamXLM_Roberta', were input into an Attention BiLSTM classifier with an input size of 768, a hidden size of 512, and 3 LSTM layers. The attention mechanism captured critical features and dependencies in multilingual sequences.

Dropout (0.5) and layer normalization were applied to stabilize training and reduce overfitting. The AdamW optimizer with a learning rate of 1×10^{-5} was used, with early stopping based on validation loss and macro F1-score ensuring robust performance.

This transliteration-aware MLM fine-tuning and Attention BiLSTM setup effectively handled transliterated and mixed-script Malayalam text.

Label	Precision	Recall	F1-Score	Support
FALSE	0.67	0.83	0.74	100
HALF TRUE	0.48	0.30	0.37	37
PARTLY FALSE	1.00	0.57	0.73	7
MOSTLY FALSE	0.50	0.45	0.47	56
Accuracy	-	-	0.61	200
Macro Avg	0.66	0.54	0.58	200
Weighted Avg	0.60	0.61	0.60	200

Table 2: Classification Report on the Test Set for Fake News Detection

Team Name	mF1	Rank
KCRL	0.6283	1
byteSizedLLM	**0.5775**	**2**
NLP_goats	0.5417	4

Table 3: Macro F1 (mF1) scores and ranks of top3 performing teams.

6 Results and Discussion

The proposed Attention BiLSTM-XLM-RoBERTa model demonstrated competitive performance in fake news detection on the Malayalam-English code-mixed dataset[3]. As shown in Table 2, the model achieved an overall accuracy of 61% with a macro F1-score of 0.58. The 'FALSE' label exhibited the highest F1-score of 0.74, while the 'HALF TRUE' label scored the lowest at 0.37, reflecting challenges posed by imbalanced data.

The fine-tuned MalayalamXLM_Roberta model, optimized with Masked Language Modeling (MLM), achieved a perplexity of 4.15, generating effective contextual embeddings. When used independently, these embeddings achieved a macro F1-score of 0.5394. Integrating them into the Attention BiLSTM classifier improved performance to a macro F1-score of 0.5775 with an optimal configuration of a learning rate of 1×10^{-5}, an LSTM hidden size of 512, and 3 LSTM layers. Other configurations, such as a learning rate of 2×10^{-5} with 256 hidden units and 2 LSTM layers, resulted in a slightly lower F1-score of 0.5718. Comparatively, an advanced encoder-decoder transformer model achieved a macro F1-score of 0.5532, reaffirming the efficiency of the Attention BiLSTM approach for small datasets.

As shown in Table 3, our team, **ByteSizedLLM**, secured second and third ranks in the shared task with macro F1-scores of 0.5775 and 0.5718, respectively. Despite most of the training data being monolingual, the multilingual XLM-RoBERTa model exhibited remarkable robustness in handling code-mixed scenarios, highlighting its adaptability across diverse linguistic contexts.

[3] https://github.com/mdp0999/Fake-News-Detection/blob/main/task2.ipynb

7 Limitations and Future Work

The model's performance was limited by the dataset size, which was restricted to 340MB of code-mixed text due to computational constraints. Additionally, inaccuracies in the transliteration process may have impacted the quality of embeddings. The imbalanced label distribution also posed challenges, particularly for minority classes like 'HALF TRUE' and 'PARTLY FALSE'.

Future work aims to overcome limitations by using larger datasets, improving transliteration, and exploring advanced architectures for better fake news detection in multilingual and code-mixed contexts.

8 Conclusion

This study proposed an Attention BiLSTM-XLM-RoBERTa model for fake news detection in Malayalam datasets. By fine-tuning XLM-RoBERTa with MLM on transliteration-aware data and integrating the embeddings into an attention-enhanced BiLSTM architecture, the approach effectively addressed linguistic and script challenges in Malayalam text. The model achieved a macro F1-score of 0.5775, securing top rankings in the shared task and demonstrating its robustness in resource-constrained settings.

Despite the predominantly monolingual nature of the training data and transliteration limitations, the model performed strongly, showcasing the ability of multilingual XLM-RoBERTa embeddings to handle diverse script variations. These results underscore the potential of multilingual models for low-resource language tasks.

References

Abhinaba Bala and Parameswari Krishnamurthy. 2023. AbhiPaw@DravidianLangTech: Multimodal abusive language detection and sentiment analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 140–146, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Pedro Henrique Arruda Faustini and Thiago Ferreira Covões. 2020. Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, 158:113503.

Gereme, Fantahun and Zhu, William and Ayall, Tewodros and Alemu, Dagmawi. 2021. Combating fake news in "low-resource" languages: Amharic fake news detection accompanied by resource crafting. *Information*, 12(1).

Mohammad Hadi Goldani, Saeedeh Momtazi, and Reza Safabakhsh. 2021. Detecting fake news with capsule neural networks. *Applied Soft Computing*, 101:106991.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ashfia Jannat Keya, Md. Anwar Hussen Wadud, M. F. Mridha, Mohammed Alatiyyah, and Md. Abdul Hamid. 2022. AugFake-BERT: Handling Imbalance through Augmentation of Fake News Using BERT to Enhance the Performance of Fake News Classification. *Applied Sciences*, 12(17).

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Zhipeng Luo and Jiahui Wang. 2023. DeepBlueAI@DravidianLangTech-RANLP 2023. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 171–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024.

CUET_NLP_GoodFellows@DravidianLangTech EACL2024: A transformer-based approach for detecting fake news in Dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192, St. Julian's, Malta. Association for Computational Linguistics.

Tanzim Rahman, Abu Raihan, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@DravidianLangTech EACL2024: Fake news classification using Malayalam-BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 223–228, St. Julian's, Malta. Association for Computational Linguistics.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in Dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Malliga S, Bharathi Raja Chakravarthi, Kogilavani S V, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, and Muskaan Singh. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 59–63, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Masood Hamed Saghayan, Seyedeh Fatemeh Ebrahimi, and Mohammad Bahrani. 2021. Exploring the Impact of Machine Translation on Fake News Detection: A Case Study on Persian Tweets about COVID-19. In *2021 29th Iranian Conference on Electrical Engineering (ICEE)*, pages 540–544.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Premjith B, Vanaja K, Mithunja S, Devika K, Hariprasath S.b, Haripriya B, and Vigneshwar E. 2024a. Overview of the second shared task on fake news detection in Dravidian languages: DravidianLangTech@EACL 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78, St. Julian's, Malta. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024b. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the Shared Task on Fake News Detection from Social Media Text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_Punctuators@DravidianLangTech-EACL2024: Transformer-based approach for detection and classification of fake news in Malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186, St. Julian's, Malta. Association for Computational Linguistics.

Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking. *Preprint*, arXiv:2011.13253.

byteSizedLLM@DravidianLangTech 2025: Multimodal Hate Speech Detection in Malayalam Using Attention-Driven BiLSTM, Malayalam-Topic-BERT, and Fine-Tuned Wav2Vec 2.0

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Daniel Iglesias
Digi Sapiens
Frankfurt, Germany
diglesias@web.de

Abstract

This research presents a robust multimodal framework for hate speech detection in Malayalam, combining fine-tuned Wav2Vec 2.0, Malayalam-Doc-Topic-BERT, and an Attention-Driven BiLSTM architecture. The proposed approach effectively integrates acoustic and textual features, achieving a macro F1-score of 0.84 on the Malayalam test set. Fine-tuning Wav2Vec 2.0 on Malayalam speech data and leveraging Malayalam-Doc-Topic-BERT significantly improved performance over prior methods using openly available models. The results highlight the potential of language-specific models and advanced multimodal fusion techniques for addressing nuanced hate speech categories, setting the stage for future work on Dravidian languages like Tamil and Telugu.

1 Introduction

Social media platforms have revolutionized digital communication, enabling the seamless exchange of multimodal content, including text, images, videos, and audio. However, the increasing prevalence of hate speech on these platforms presents significant challenges for content moderation. The detection of such content requires advanced multimodal analysis techniques that effectively integrate textual and speech features to capture intent and context.

The Shared Task on Multimodal Hate Speech Detection in Dravidian Languages (DravidianLangTech@NAACL 2025) (Lal G et al., 2025), part of the Multimodal Social Media Data Analysis (MSMDA) initiative, promotes research in analyzing complex social media data using multimodal approaches. The MSMDA-DL shared task focuses on Tamil and Malayalam, two linguistically rich Dravidian languages, emphasizing the need for innovative multimodal natural language processing (NLP) techniques. This study specifically addresses **Task 1: Multimodal Hate Speech Detection in Malayalam**, where detecting hate speech requires integrating textual embeddings and acoustic features to distinguish between non-hate and various hate categories effectively.

This paper presents a multimodal classification approach combining Attention-Driven BiLSTM, BERT-Base, and Wav2Vec models to enhance Malayalam hate speech detection. The proposed architecture captures semantic and phonetic nuances, leveraging BERT-Base for text representation and Wav2Vec for speech-based feature extraction. The study provides insights into data preprocessing, model architecture, and classification performance, contributing to the broader understanding of multimodal hate speech detection in low-resource languages.

This study details our data preprocessing, Wav2Vec fine-tuning, and multimodal classifier design, introducing optimizations that improve detection accuracy and scalability. Our results provide insights into the challenges of multimodal NLP, contributing to advancements in hate speech detection for low-resource languages.

2 Related Work

Multimodal approaches for analyzing social media data have advanced significantly, particularly for underrepresented languages like Tamil, Telugu, and Malayalam. Banerjee et al. (2020) used an autoregressive XLNet for sentiment analysis on Tamil-English and Malayalam-English datasets, highlighting the challenges of multilingual and code-mixed data.

B et al. (2022) introduced the DravidianMultiModality Dataset, incorporating textual, audio, and visual features from product and movie review videos, underscoring the benefits of multimodal sentiment analysis.

Similarly, B et al. (2023) applied multimodal deep learning to disaster response, demonstrating

how text and image integration aids real-world classification tasks.

The DravidianLangTech 2024 shared tasks (B et al., 2024; Premjith et al., 2024b,a) advanced multilingual and multimodal research, focusing on sentiment analysis, hate speech detection, and language identification for Dravidian languages. These initiatives foster innovation in handling the linguistic and cultural diversity of Tamil and Malayalam.

Building on this, the MSMDA shared task in Malayalam integrates textual and audio features to enhance hate speech detection. This effort tackles challenges in code-mixed content, complex morphology, and rich phonetic structures, pushing research forward in multimodal NLP for underrepresented languages.

3 Dataset

3.1 Fine-tuning Wav2Vec for Malayalam Speech Recognition

To enhance the performance of our multimodal hate speech detection model, we fine-tuned Wav2Vec 2.0 using a Malayalam speech recognition dataset sourced from the ULCA-ASR dataset corpus[1]. This 637.88-hour unlabelled Malayalam speech dataset supports fine-tuning Wav2Vec 2.0 base, enhancing phonetic and acoustic modeling for improved speech feature extraction in downstream tasks.

3.2 Hate Speech Dataset for Multimodal Testing

The Malayalam hate speech dataset, collected from YouTube, includes 933 utterances labeled into five categories: Non-Hate (**N**) and four hate subcategories—Gender (**G**), Political (**P**), Religious (**R**), and Personal Defamation (**C**) (Sreelakshmi et al., 2024). Each sample contains both audio and text for comprehensive multimodal analysis.

The dataset is split into 883 training samples (794 train, 89 dev) and 50 test samples. This structured labeling aids in effective classification and highlights hate speech characteristics in Malayalam.

Label	Train	Test	Total
N	406	10	416
C	186	10	196
P	118	10	128
R	91	10	101
G	82	10	92
Total	883	50	933

Table 1: Label distribution for Malayalam hate speech dataset.

4 Methodology

This section presents the methodology for multimodal hate speech detection in Malayalam, integrating fine-tuned Wav2Vec 2.0 for speech, Malayalam-Doc-Topic-BERT for textual embeddings, and an attention-driven BiLSTM-BERT-Wav2Vec classifier for fusion.

4.1 Fine-tuning Wav2Vec 2.0 for Malayalam Speech Recognition

The Wav2Vec 2.0 base model was fine-tuned on the ULCA-ASR dataset corpus of unlabelled Malayalam speech. The fine-tuning process utilized Facebook's Fairseq framework, optimizing the model to capture phonetic and acoustic nuances specific to Malayalam. This adaptation allowed the model to generate robust speech embeddings for downstream tasks, addressing the rich phonetic diversity and morphological complexity of Malayalam. The fine-tuned model serves as the foundation for extracting audio features in the proposed multimodal approach.

4.2 Malayalam-Doc-Topic-BERT

For textual embeddings, we selected the IndicSBERT model, l3cube-pune malayalam-sentence-bert-nli[2] (Mirashi et al., 2024), which was further fine-tuned on the L3Cube-IndicNews Corpus. This corpus encompasses three sub-datasets: Long Document Classification (LDC), Long Paragraph Classification (LPC), and Short Headline Classification (SHC), representing different document lengths. By training on a combination of these datasets, the Malayalam-Doc-Topic-BERT model achieves consistent performance across varied text lengths. It captures contextual semantics and document-level information for Malayalam hate speech detection.

4.3 Attention-Driven BiLSTM-BERT-Wav2Vec Classifier

This study presents a hybrid Attention-Driven BiLSTM-BERT-Wav2Vec model (Liu and Guo,

[1] https://github.com/Open-Speech-EkStep/ULCA-asr-dataset-corpus

[2] https://huggingface.co/l3cube-pune/malayalam-topic-all-doc

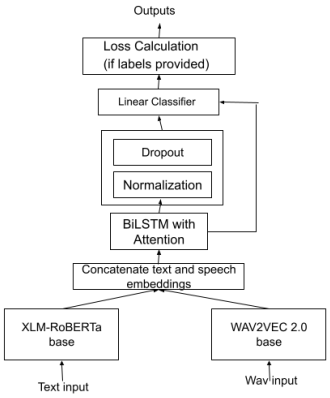


Figure 1: Architecture of the Attention-Driven BiLSTM-XLM-RoBERTa-Wav2Vec Classifier.

2019; Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Kodali et al., 2025; Manukonda and Kodali, 2025, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b) for multimodal classification, integrating text and speech features.

Text input is processed via a fine-tuned Malayalam-Doc-Topic-BERT, generating contextual embeddings:

$$\mathbf{X}_t = \textbf{BERT}(input\_ids, attention\_mask) \quad (1)$$

Speech input is handled by a fine-tuned Wav2Vec model, producing acoustic embeddings:

$$\mathbf{X}_s = \textbf{Wav2Vec}(audio\_features) \quad (2)$$

Both embeddings are concatenated:

$$\mathbf{X} = [\mathbf{X}_t; \mathbf{X}_s] \quad (3)$$

A BiLSTM extracts temporal patterns:

$$\mathbf{H}_t = [\mathbf{H}_{fwd,t}; \mathbf{H}_{bwd,t}] \quad (4)$$

An attention mechanism assigns weights α_t to focus on key features:

$$\alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T} \exp(\mathbf{a}_t)}, \quad \mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t \quad (5)$$

Residual components, including layer normalization and dropout, enhance generalization, robustness, and stabilize training.

$$\mathbf{H}_{dropout} = Dropout(LayerNorm(\mathbf{H}_{attended})) \quad (6)$$

A fully connected layer maps features to classification logits:

$$\textbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \quad (7)$$

The model is optimized using **cross-entropy loss**:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (8)$$

This hybrid architecture (Figure 1) effectively integrates linguistic and acoustic insights, leveraging BERT, Wav2Vec, and BiLSTM with attention to enhance multimodal hate speech detection in Malayalam.

5 Experiment Setup

Our approach comprises fine-tuning Wav2Vec 2.0 base on Malayalam speech and developing a multimodal hate speech detection framework. The Wav2Vec 2.0 base model was fine-tuned on the ULCA-ASR Malayalam dataset (637.88 hours of unlabelled speech), with preprocessing steps including resampling to 16 kHz and noise reduction. Training used Fairseq, a tri-stage learning rate schedule, and ran for 50 epochs with Adam (lr = 1e-4). The best checkpoint, based on Word Error Rate (WER), was used for speech embeddings.

A hybrid model combined BERT base text embeddings with Wav2Vec base speech embeddings, processed through a BiLSTM (512 hidden units, 2 layers) with attention. Dropout (0.3) and layer normalization ensured stability. The final classifier predicted one of five labels.

Training employed PyTorch, AdamW (lr = 2e-5), and a ReduceLROnPlateau scheduler, running for 10 epochs on GPU, saving the best macro F1-score checkpoint. Evaluation measured accuracy, precision, recall, and F1-score, confirming the effectiveness of Wav2Vec-BERT fusion with BiLSTM and attention for Malayalam hate speech detection.

The model was assessed using the macro F1-score per DravidianLangTech guidelines, with scikit-learn generating precision, recall, and F1-score for all categories.

6 Results and Discussion

The fine-tuned Wav2Vec 2.0 model on the ULCA-ASR Malayalam dataset achieved a WER of 17.4%, enhancing phonetic representation for multimodal classification. Our Attention-Driven BiLSTM-BERT-Wav2Vec model attained an accuracy of

Label	Precision	Recall	F1-score	Support
C	0.77	1.00	0.87	10
G	1.00	0.70	0.82	10
N	0.83	1.00	0.91	10
P	0.88	0.70	0.78	10
R	0.80	0.80	0.80	10
Accuracy	-	-	0.84	50
Macro Avg	0.86	0.84	0.84	50
Weighted Avg	0.86	0.84	0.84	50

Table 2: Classification Report on the Test Set for Multimodal Hate Speech Detection, including precision, recall, F1-score, and support for each label.

84%, with macro and weighted F1-scores of 0.84. The best performance was observed in **C** (0.87) and **N** (0.91) categories, while **G** (0.82) showed high precision but lower recall. The **P** and **R** categories had F1-scores of 0.78 and 0.80, indicating challenges in detecting implicit hate speech.

Table 3 compares our model with top teams in DravidianLangTech@NAACL 2025. Our model outperformed all tested architectures, achieving a macro F1-score of 0.8360 due to effective integration of Malayalam-Doc-Topic-BERT and Wav2Vec 2.0 embeddings with BiLSTM and attention mechanisms. Training code and evaluation scripts are publicly available on GitHub[3], ensuring reproducibility.

Team Name	mF1	Rank
SSNTrio	0.7511	1
lowes	0.7367	2
MNLP	0.6135	3
byteSizedLLM	**0.5831**	**4**
KEC_Tech_Titans	0.5114	5
Attention-BiLSTM-BERT-Wav2Vec: 0.8360		

Table 3: Performance comparison in Multimodal Hate Speech Detection at DravidianLangTech@NAACL 2025.

Our submission ranked 4th (macro F1-score: 0.5831) as fine-tuning was incomplete at the deadline, requiring the use of an open-source Wav2Vec 2.0 base and XLM-RoBERTa base models. The `Malayalam-Doc-Topic-BERT` replacement improved performance. The top team, **SSNTrio**, achieved 0.7511.

Future work includes extending this approach to Telugu and Tamil, improving fusion techniques like hierarchical attention, and mitigating dataset imbalances.

[3] https://github.com/mdp0999/Multimodal-Hate-Speech-in-Malayalam

7 Limitations and Future Work

Despite strong overall performance, our model has several limitations. First, it struggles with implicit hate speech, particularly in the **P** (Political) and **R** (Religious) categories, where nuanced language reduces recall. Second, reliance on pre-trained multilingual models limits adaptability to low-resource languages. Third, dataset imbalances affect recall, as seen in the **G** (Gender) category, which had high precision (1.00) but low recall (0.70), indicating missed instances of gender-based hate speech. Fourth, fine-tuning of Wav2Vec 2.0 was incomplete at submission, impacting final performance. Future work will focus on language-specific fine-tuning, dataset expansion, and improved multimodal fusion techniques to mitigate these limitations.

8 Conclusion

This work presents a novel multimodal framework combining Attention-Driven BiLSTM, fine-tuned Wav2Vec 2.0, and Malayalam-Doc-Topic-BERT for hate speech detection in Malayalam, achieving a macro F1-score of 0.84 and surpassing existing baselines in performance. The proposed method effectively integrates acoustic and textual features, demonstrating its ability to address the linguistic and cultural complexities of Malayalam. The use of fine-tuned Wav2Vec 2.0 and Malayalam-Doc-Topic-BERT emphasizes the importance of tailored, language-specific models for resource-scarce languages.

References

Premjith B, Bharathi Raja Chakravarthi, Malliga Subramanian, Bharathi B, Soman Kp, Dhanalakshmi V, Sreelakshmi K, Arunaggiri Pandian, and Prasanna Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 254–260, Dublin, Ireland. Association for Computational Linguistics.

Premjith B, Jyothish G, Sowmya V, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, Bharathi B, Saranya Rajiakodi, Rahul Ponnusamy, Jayanth Mohan, and Mekapati Reddy. 2024. Findings of the shared task on multimodal social media data analysis in Dravidian languages (MSMDA-DL)@DravidianLangTech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61, St. Julian's, Malta. Association for Computational Linguistics.

Premjith B, Jyothish Lal G, Sowmya V, Bharathi Raja Chakravarthi, Rajeswari Natarajan, Nandhini K, Abirami Murugappan, Bharathi B, Kaushik M, Prasanth Sn, Aswin Raj R, and Vijai Simmon S. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in Tamil and Malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Shubhanker Banerjee, Arun Jayapal, and Sajeetha Thavareesan. 2020. Nuig-shubhanker@dravidian-codemix-fire2020: Sentiment analysis of code-mixed dravidian text using xlnet. *Preprint*, arXiv:2010.07773.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Aishwarya Mirashi, Srushti Sonavane, Purva Lingayat, Tejas Padhiyar, and Raviraj Joshi. 2024. L3cube-indicnews: News-based short text and long document classification datasets in indic languages. *arXiv preprint arXiv:2401.02254*.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi,

Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

byteSizedLLM@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Dravidian Languages Using XLM-RoBERTa and Attention-BiLSTM

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Maharajan Pannakkaran
ASRlytics
Hyderabad, India
mahamca.kovai@gmail.com

Abstract

This study presents a hybrid model integrating TamilXLM-RoBERTa and MalayalamXLM-RoBERTa with BiLSTM and attention mechanisms to classify AI-generated and human-written product reviews in Tamil and Malayalam. The model employs a transliteration-based fine-tuning strategy, effectively handling native, Romanized, and mixed-script text. Despite being trained on a relatively small portion of data, our approach demonstrates strong performance in distinguishing AI-generated content, achieving competitive macro F1 scores in the DravidianLangTech 2025 shared task. The proposed method showcases the effectiveness of multilingual transformers and hybrid architectures in tackling low-resource language challenges.

1 Introduction

The rapid advancement of artificial intelligence (AI) has significantly transformed natural language processing (NLP) and content generation. While these developments enhance text-based applications, they also facilitate the proliferation of AI-generated content, posing challenges to domains that rely on textual authenticity, such as online product reviews. The increasing sophistication of synthetic text generation necessitates effective detection mechanisms to preserve content credibility (Ben Jabeur et al., 2023).

To address this issue, the Shared Task on Detecting AI-Generated Product Reviews in Dravidian Languages, organized as part of DravidianLangTech 2025[1], focuses on detecting synthetic content in Malayalam and Tamil (Premjith et al., 2025). While extensive research exists for high-resource languages like English, AI-generated text detection in Dravidian languages remains underexplored. The complex morphological structures, agglutinative nature, and unique syntactic properties of these languages present additional challenges.

We propose a hybrid model combining fine-tuned, transliteration-aware XLM-RoBERTa (Conneau et al., 2019) with an Attention-BiLSTM (Liu and Guo, 2019) classifier. XLM-RoBERTa captures linguistic nuances through robust cross-lingual representation learning, while the BiLSTM layer, enhanced with attention mechanisms, improves sequential dependency learning and feature prioritization. This integration of transformer-based architectures with recurrent neural networks enhances the detection of AI-generated content.

This paper details our methodology, experimental setup, and results, demonstrating the effectiveness of our approach. We also discuss the challenges of detecting AI-generated text in Dravidian languages and explore future directions for improving content authenticity verification in low-resource linguistic settings.

2 Related Work

The rise of generative AI has raised concerns about its misuse in creating deceptive content like fake product reviews. Luo et al. (2023); Ben Jabeur et al. (2023) proposed a supervised learning framework using statistical theories to detect AI-generated reviews by identifying outliers in feature distributions. Similarly, Gupta et al. (2024) reviewed advancements in fake review detection, emphasizing hybrid frameworks and challenges in detecting AI-generated content.

AI-generated reviews typically feature two categories: novel features from large language models (LLMs) and traditional linguistic features. LLM-generated text tends to be more readable but templated due to predictive word selection, while human-authored text shows more unpredictability and lexical diversity (Guo et al., 2023). Detection metrics like perplexity and burstiness, used in tools

[1] https://codalab.lisn.upsaclay.fr/competitions/20700

like GPTZero (Tian and Cui, 2023), measure text randomness and aid in identifying AI-generated content (Cai and Cui, 2023; Liang et al., 2023).

Traditional linguistic features, including sentiment polarity, adjective ratios, and reviewer behavior, have been effective in detecting fake reviews (Yin et al., 2021; Kumar et al., 2022). However, integrating LLM-based and traditional features remains underexplored.

Detecting AI-generated reviews in Malayalam and Tamil is challenging due to their complex morphology and syntax. This work addresses the gap by integrating LLM-based and linguistic features for better detection in low-resource languages.

3 Dataset

This study employs a dataset for detecting AI-generated product reviews in Tamil and Malayalam (Premjith et al., 2025). The task dataset is labeled into two categories: **AI-generated** and **HUMAN-written** reviews. The statistics for both languages are presented in Tables 1 and 2.

Label	Train	Test	Total
AI	405	48	453
HUMAN	403	52	455
Total	808	100	908

Table 1: Tamil dataset distribution across training and test splits.

Label	Train	Test	Total
HUMAN	400	100	500
AI	400	100	500
Total	800	200	1000

Table 2: Malayalam dataset distribution across training and test splits.

The Tamil dataset consists of 908 reviews, with 808 for training and 100 for testing, maintaining a balanced distribution between AI-generated and HUMAN-written reviews. Similarly, the Malayalam dataset comprises 1,000 reviews, with 800 for training and 200 for testing, equally split across both categories. Both datasets follow a **90:10 ratio** for training and development, ensuring stratified splits for robust evaluation.

4 Methodology

This study employs a hybrid Attention BiLSTM-XLM-RoBERTa model (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Kodali et al., 2025; Manukonda and Kodali, 2025, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b) to classify AI-generated and HUMAN-written product reviews in Tamil and Malayalam. The architecture, shown in Figure 1, combines the strengths of fine-tuned XLM-RoBERTa embeddings, a bidirectional LSTM (BiLSTM), and an attention mechanism to effectively extract and process features for classification.

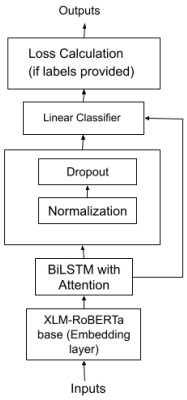


Figure 1: Architecture of the BiLSTM-XLM-RoBERTa Classifier Model.

4.1 Transliteration aware XLM-RoBERTa Fine-tuning

The TamilXLM-RoBERTa[2] and MalayalamXLM-RoBERTa[3] models were fine-tuned using a transliteration strategy with the **IndicTrans** tool (Bhat et al., 2015), leveraging approximately 300MB of monolingual text from AI4Bharath (Kunchukuttan et al., 2020) for each language. The dataset included three variations: native script text, fully transliterated text in Roman script, and partially transliterated text where 20–70% of words were transliterated. This approach enables the model to handle native scripts, Romanized text, and mixed-script text, which are common in social media communication.

4.2 Attention-BiLSTM-XLM-RoBERTa Classifier

The Attention-BiLSTM-XLM-RoBERTa classifier integrates contextual embeddings, sequential modeling, and attention-based feature selection. The

[2] https://huggingface.co/bytesizedllm/TamilXLM_Roberta

[3] https://huggingface.co/bytesizedllm/MalayalamXLM_Roberta

input sequence is processed by a fine-tuned XLM-RoBERTa model to generate contextual embeddings:

$$\mathbf{X} = \textbf{XLMRoBERTa}(input\_ids, att\_mask) \tag{1}$$

These embeddings are passed through a BiLSTM layer, capturing sequential dependencies by concatenating forward and backward hidden states:

$$\mathbf{H}_t = [\mathbf{H}_{fwd,t}; \mathbf{H}_{bwd,t}] \tag{2}$$

An attention mechanism assigns importance weights to hidden states:

$$\mathbf{a}_t = \tanh(\mathbf{W}_{att} \cdot \mathbf{H}_t), \quad \alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T} \exp(\mathbf{a}_t)} \tag{3}$$

The weighted sum of hidden states forms the attended representation:

$$\mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t \tag{4}$$

Layer normalization and dropout stabilize training:

$$\mathbf{H}_{dropout} = Dropout(LayerNorm(\mathbf{H}_{attended})) \tag{5}$$

Finally, a classification layer produces logits:

$$\textbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \tag{6}$$

Training is optimized using the cross-entropy loss function:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{7}$$

This architecture effectively combines XLM-RoBERTa embeddings, BiLSTM for sequential learning, and attention for key feature selection, enhancing multi-label classification performance.

5 Experiment Setup

We fine-tuned Tamil XLM-RoBERTa and Malayalam XLM-RoBERTa for monolingual and multilingual text classification. The datasets were processed using a data preprocessing pipeline, and labels were encoded as integers for multi-class classification. The data was split into 90% training and 10% validation using a stratified approach.

The fine-tuned XLM-RoBERTa embeddings were integrated with a BiLSTM layer with a hidden size of 512, 3 LSTM layers, and a dropout probability of 0.3. An attention mechanism was added to refine the feature representation. The model was trained for 10 epochs using the AdamW optimizer with a learning rate of 2.5×10^{-5}, weight decay of 0.01, and a linear learning rate scheduler. A batch size of 16 was used, and gradient clipping with a maximum norm of 1.0 was applied for stability.

Validation used accuracy and macro F1-score per epoch, saving the best model for each language to ensure effective fine-tuning for detecting AI-generated reviews in Tamil and Malayalam.

6 Results and Discussion

Team Name	mF1	Rank
KaamKro	0.9199	1
Nitiz - StarAtNyte	0.915	2
Three_Musketeers	0.915	2
SSNTrio	0.9147	3
byteSizedLLM	**0.9**	**4**
Lowes	0.9	4

Table 3: Macro F1 (mF1) scores and ranks of the top 4 performing teams on the Malayalam test set.

Team Name	mF1	Rank
KEC_AI_NLP	0.97	1
CUET_NLP_FiniteInfinity	0.97	1
CIC-NLP	0.96	2
KaamKro	0.95	3
KEC-Elite-Analysts	0.9499	4
byteSizedLLM	**0.94**	**5**

Table 4: Macro F1 (mF1) scores and ranks of the top 5 performing teams on the Tamil test set.

Our experiments demonstrate the effectiveness of the fine-tuned TamilXLM-RoBERTa and MalayalamXLM-RoBERTa models in classifying AI-generated and HUMAN-written product reviews[4]. The perplexity scores achieved by the models underline their capability to adapt to the linguistic nuances of the respective languages, with the Malayalam model achieving a perplexity of 4.1 and the Tamil model achieving a perplexity of 4.9.

Table 3 highlights the performance of the top-performing teams on the Malayalam test set. Our

[4] https://github.com/mdp0999/Detecting-AI-generated-product-reviews

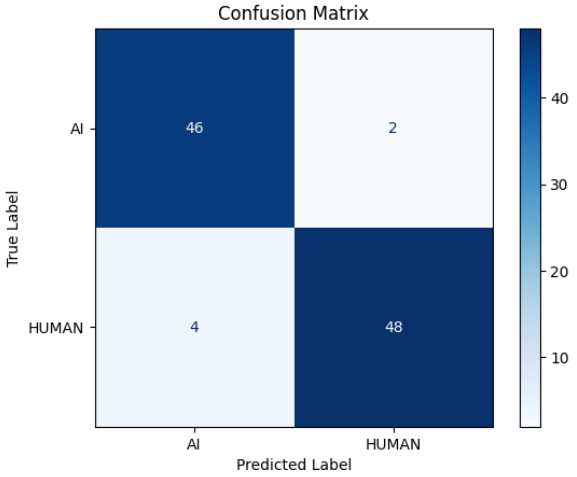


Figure 2: Confusion Matrix for Tamil AI-Generated vs. Human-Written Review Classification

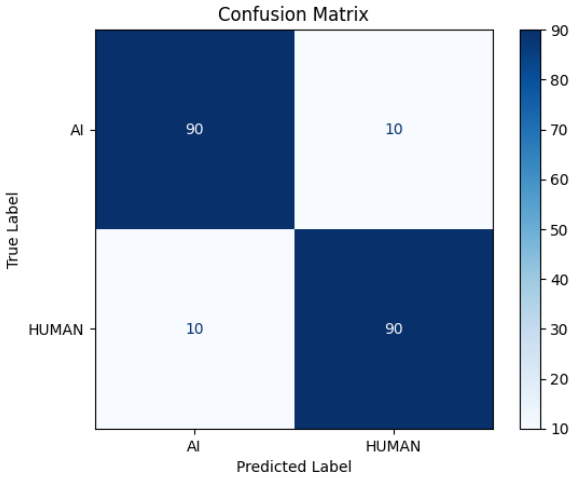


Figure 3: Confusion Matrix for Malayalam AI-Generated vs. Human-Written Review Classification

team, **byteSizedLLM**, secured a shared **4th place** with a Macro F1 (mF1) score of **0.9**. This outcome reflects the strength of our hybrid architecture, which integrates fine-tuned XLM-RoBERTa embeddings with BiLSTM layers and attention mechanisms to address the complexities of Malayalam text effectively.

For the Tamil test set, as summarized in Table 4, our team achieved an mF1 score of **0.94**, placing **5th** among the top teams. The slightly higher perplexity for Tamil indicates challenges in modeling the language, potentially due to its linguistic structure or the dataset's characteristics. Nonetheless, the results validate the robustness of our transliteration-based fine-tuning strategy in managing native, Romanized, and mixed-script text.

The confusion matrices reveal that the model achieves balanced performance across both classes, with very few false positives and false negatives. However, the slightly lower recall for the HUMAN class in Tamil suggests that the model may occasionally misclassify human-written reviews as AI-generated, warranting further optimization. For better understanding, please refer to Fig.2 for Tamil and Fig.3 for Malayalam.

6.1 Limitations and Future Work

Our models were fine-tuned on a limited portion of the available datasets (approximately 300MB per language), constrained by computational resources. This limited dataset size may have restricted the models' ability to fully exploit the linguistic diversity of Tamil and Malayalam. Despite these constraints, the models demonstrated strong performance, but further improvements could be achieved with larger datasets and enhanced computational capabilities.

Future work will focus on scaling the fine-tuning process to utilize more extensive datasets, enabling deeper language modeling. Additionally, adopting advanced strategies such as dynamic data augmentation, multi-task learning, and incorporating more sophisticated preprocessing techniques could further refine model performance. These enhancements aim to reduce perplexity and boost classification accuracy for AI-generated product reviews across multilingual contexts.

7 Conclusion

This study successfully fine-tuned TamilXLM-RoBERTa and MalayalamXLM-RoBERTa models to classify AI-generated and HUMAN-written product reviews. Despite computational constraints limiting the dataset size, the models delivered strong performance, achieving Macro F1 scores of **0.94** for Tamil and **0.9** for Malayalam, ranking among the top teams in their respective tasks. The transliteration-based fine-tuning strategy, combined with a robust hybrid architecture, proved effective in processing diverse scripts, including native, Romanized, and mixed-script text. Remarkably, although the training data was monolingual, the approach demonstrated an ability to generalize to multilingual and mixed-script scenarios, making it highly adaptable for real-world multilingual text classification challenges.

References

Sami Ben Jabeur, Hossein Ballouk, Wissal Ben Arfi, and Jean-Michel Sahut. 2023. Artificial intelligence applications in fake review detection: Bibliometric analysis and future avenues for research. *Journal of Business Research*, 158:113631.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Shuyang Cai and Wanyun Cui. 2023. Evade chatgpt detectors via a single space. *Preprint*, arXiv:2307.02599.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

Richa Gupta, Vinita Jindal, and Indu Kashyap. 2024. Recent state-of-the-art of fake review detection: a comprehensive review. *The Knowledge Engineering Review*, 39:e8.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Ajay Kumar, Ram Gopal, Ravi Shankar, and Kim Tan. 2022. Fraudulent review detection model focusing on emotional expressions and explicit aspects: investigating the potential of feature engineering. *Decision Support Systems*, 155:113728.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Preprint*, arXiv:2304.02819.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Jiwei Luo, Jian Luo, Guofang Nan, and Dahui Li. 2023. Fake review detection system for online e-commerce platforms: A supervised general mixed probability approach. *Decision Support Systems*, 175:114045.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Edward Tian and Alexander Cui. 2023. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods".

Chunyong Yin, Haoqi Cuan, Yuhang Zhu, and Zhichao Yin. 2021. Improved fake reviews detection model based on vertical ensemble tri-training and active learning. *ACM Trans. Intell. Syst. Technol.*, 12:33:1–33:19.

byteSizedLLM@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media Using XLM-RoBERTa and Attention-BiLSTM

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Maharajan Pannakkaran
ASRlytics
Hyderabad, India
mahamca.kovai@gmail.com

Abstract

This research investigates abusive comment detection in Tamil and Malayalam, focusing on code-mixed, multilingual social media text. A hybrid Attention BiLSTM-XLM-RoBERTa model was utilized, combining fine-tuned embeddings, sequential dependencies, and attention mechanisms. Despite computational constraints limiting fine-tuning to a subset of the AI4Bharath dataset, the model achieved competitive macro F1-scores, ranking 6th for both Tamil and Malayalam datasets with minor performance differences. The results emphasize the potential of multilingual transformers and the need for further advancements, particularly in addressing linguistic diversity, transliteration complexity, and computational limitations.

1 Introduction

Social media platforms enable communication but are increasingly misused for abuse and harassment. Women often face hateful and threatening comments, reflecting deep-rooted societal biases. This gender-based cyberbullying leads to severe psychological, social, and professional harm. Addressing this issue is vital for creating safer online environments.

The Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media at DravidianLangTech@NAACL 2025[1] seeks to tackle this pressing issue by advancing research on online content moderation. This task focuses on detecting abusive comments targeting women in Tamil and Malayalam, two Dravidian languages predominantly spoken in South India. As low-resource languages in the field of natural language processing (NLP), Tamil and Malayalam present unique challenges for developing robust machine learning models. Furthermore, identifying abusive content in these languages is crucial for empowering marginalized communities and bridging the linguistic gap in content moderation research.

We propose a hybrid model that integrates fine-tuned Tamil and Malayalam XLM-RoBERTa, optimized for transliteration-aware data, with an Attention-BiLSTM classifier. XLM-RoBERTa captures cross-lingual and contextual representations, handling mixed-script inputs, while the Attention-BiLSTM identifies sequential dependencies and highlights key features. This fusion combines transformer-based embeddings and recurrent architectures for effective abuse detection in low-resource languages.

This paper details our methodology, experiments, and results, demonstrating our model's effectiveness in detecting abusive comments. We also discuss challenges encountered and suggest future directions for abusive language detection in low-resource settings.

2 Related Work

Research on detecting Hate, Offensive, and Abusive Speech in CodeMix Dravidian languages like Kannada-English, Malayalam-English, and Tamil-English has grown recently. However, challenges such as linguistic diversity, complex grammar, polysemous words, and limited annotated data persist (Anbukkarasi and Varadhaganapathy, 2023; Chakravarthi et al., 2021c). Shared tasks like DravidianLangTech 2021 and HASOC-DravidianCodeMix (Chakravarthi et al., 2021a,b), alongside annotated datasets (Chakravarthi et al., 2020, 2021b, 2022; Devi, 2021; Jose et al., 2020), have enabled significant advancements in this domain.

Participating teams in these shared tasks have utilized multilingual pre-trained transformers for their contextual understanding and fine-tuning capabilities. For example, Saha et al. (2021) leveraged models like XLM-RoBERTa-large, MuRIL, and In-

[1] https://codalab.lisn.upsaclay.fr/competitions/20701

dicBERT, while Balouchzahi et al. (2021) proposed the COOLI Ensemble model with CountVectors and classifiers such as MLP and XGBoost. Other approaches include handling class imbalance with innovative loss functions (Tula et al., 2021; Vasantharajan and Thayasivam, 2021) and employing strategies like pseudo-labeling, multi-task learning, and selective translation for fine-tuning (Hande et al., 2021a,b; Vasantharajan and Thayasivam, 2021).

Traditional machine learning methods, such as SVMs and Random Forest, have also been explored with feature extraction techniques like TF-IDF (Sivalingam and Thavareesan, 2021). Indic-specific models like IndicBART (Dabre et al., 2022) have shown potential in tasks like translation and summarization. Despite these efforts, there remains no widely recognized pre-trained model for hate speech detection in CodeMix Dravidian languages.

Abusive comment detection in Tamil was a key focus of DravidianLangTech 2022(Priyadharshini et al., 2022), where datasets highlighted challenges in handling code-mixed Tamil-English text. In DravidianLangTech 2023(Bala and Krishnamurthy, 2023), the scope expanded to include both Tamil and Telugu, offering new datasets and benchmarks. These tasks spurred advancements in multilingual transformers, ensemble learning, and strategies for tackling class imbalance and data scarcity, further enriching abusive comment detection research in Dravidian languages.

Despite recent advancements, there remains significant scope to improve existing models and develop new, robust approaches for abusive comment detection in Dravidian languages. Addressing challenges like linguistic diversity, complex code-mixing, and limited annotated data will be critical to advancing this field further

3 Dataset

The goal of this task is to identify whether a given comment contains abusive content or not. The task dataset comprises text in Tamil and Malayalam(Priyadharshini et al., 2022), two low-resource languages spoken in South India. Each comment is annotated with binary labels: **Abusive** and **Non-Abusive**.

3.1 Malayalam Dataset

The Malayalam dataset contains a total of 3562 comments, divided into **2933** training and **629** testing instances. Table 1 provides a detailed breakdown of the dataset statistics for Malayalam.

Label	Train	Test	Total
Abusive	1531	323	1854
Non-Abusive	1402	306	1708
Total	2933	629	3562

Table 1: Dataset statistics for Malayalam, including total counts for each label and split.

3.2 Tamil Dataset

The Tamil dataset contains a total of 3388 comments, with **2790** for training and **598** for testing. Table 2 provides a detailed breakdown of the dataset statistics for Tamil.

Label	Train	Test	Total
Non-Abusive	1424	305	1729
Abusive	1366	293	1659
Total	2790	598	3388

Table 2: Dataset statistics for Tamil, including total counts for each label and split.

4 Methodology

This study employs a hybrid Attention BiLSTM-XLM-RoBERTa model(Liu and Guo, 2019; Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Conneau et al., 2019; Manukonda and Kodali, 2025; Kodali et al., 2025; Manukonda and Kodali, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b) to classify abusive and non-abusive comments in Tamil and Malayalam. The architecture, shown in Figure 1, combines fine-tuned XLM-RoBERTa embeddings, a bidirectional LSTM (BiLSTM), and an attention mechanism to effectively extract and process features for binary classification.

4.1 Transliteration aware XLM-RoBERTa Fine-tuning

The TamilXLM-RoBERTa and MalayalamXLM-RoBERTa models were fine-tuned using approximately 300MB of monolingual text from AI4Bharath[2](Kunchukuttan et al., 2020) for each language. To handle the diverse script usage in comments, the **IndicTrans**(Bhat et al., 2015) transliteration tool was employed to create three

[2]https://github.com/AI4Bharat/indicnlp_corpus

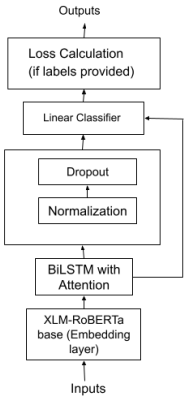


Figure 1: Architecture of the BiLSTM-XLM-RoBERTa Classifier Model.

variations of the dataset: native script text, fully transliterated text in Roman script, and partially transliterated text (20–70% of words transliterated). This approach ensures compatibility with native scripts, Romanized text, and mixed-script text, which are prevalent in social media communication.

4.2 Attention-BiLSTM-XLM-RoBERTa Classifier

The Attention-BiLSTM-XLM-RoBERTa classifier combines three key components: contextual understanding, sequential learning, and attention-based feature selection. XLM-RoBERTa generates contextual embeddings, which are refined by a BiLSTM layer to capture sequential dependencies. An attention mechanism identifies and emphasizes important features, enhancing interpretability. Residual layer normalization and dropout are applied for stability. Finally, a classification layer produces logits, optimized using cross-entropy loss. This hybrid model effectively detects abusive comments in Tamil and Malayalam.

5 Experiment Setup

We fine-tuned TamilXLM-RoBERTa[3] and MalayalamXLM-RoBERTa[4] for multilingual, code-mixed text. Data preprocessing included removing punctuation, HTML tags, and noise, with labels encoded for binary classification. A 90:10 stratified split ensured balanced training and validation sets.

[3] https://huggingface.co/bytesizedllm/TamilXLM_Roberta

[4] https://huggingface.co/bytesizedllm/MalayalamXLM_Roberta

Using **IndicTrans**, three dataset variations were created: native script, fully Romanized, and partially transliterated text, enabling the model to handle mixed-script social media text. Fine-tuned XLM-RoBERTa embeddings were combined with a BiLSTM layer for sequential learning and an attention mechanism for refined feature representation.

The model was trained for 10 epochs with AdamW (2×10^{-5} learning rate, 0.01 weight decay) and a linear scheduler. Gradient clipping (max norm 1.0) stabilized training. Validation after each epoch used accuracy and macro F1-score, with the best model per language selected based on macro F1-score, ensuring robust fine-tuning for detecting abusive comments in Tamil and Malayalam.

Team Name	mF1	Rank
CUET_Agile	0.7883	1
MSM_CUET	0.7873	2
Incepto	0.7864	3
Lowes	0.7824	4
Necto	0.7821	5
byteSizedLLM	**0.7820**	**6**

Table 3: Macro F1 (mF1) scores and ranks of the top 6 performing teams on the Tamil test set.

Team Name	mF1	Rank
Habiba A, G Agila	0.7571	1
CUET_Agile	0.7234	2
CUET_Novice	0.7083	3
Incepto	0.7058	4
Lowes	0.7001	5
byteSizedLLM	**0.6964**	**6**

Table 4: Macro F1 (mF1) scores and ranks of the top 6 performing teams on the Malayalam test set.

6 Results and Discussion

The fine-tuned multilingual model achieved a perplexity of 4.9 for Tamil and 4.1 for Malayalam, demonstrating effective language modeling performance across both languages.

The model performed better on the Tamil dataset, achieving an accuracy of 78% with F1-scores of 0.79 for Abusive and 0.77 for Non-Abusive content. In contrast, the Malayalam dataset had a lower accuracy of 70%, with F1-scores of 0.67 for Abusive and 0.73 for Non-Abusive content. The Tamil dataset exhibited a more balanced precision (0.80)

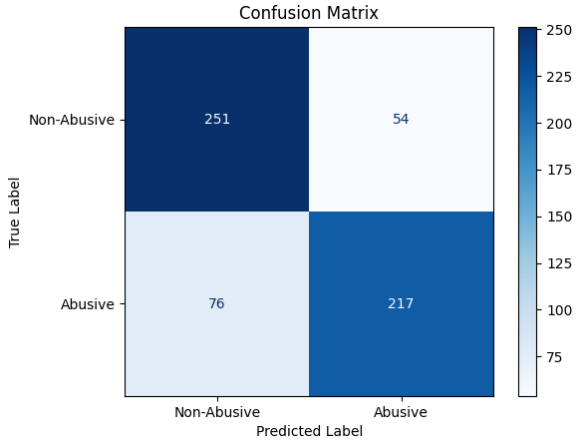


Figure 2: Confusion Matrix for Abusive and Non-Abusive Classification in Tamil

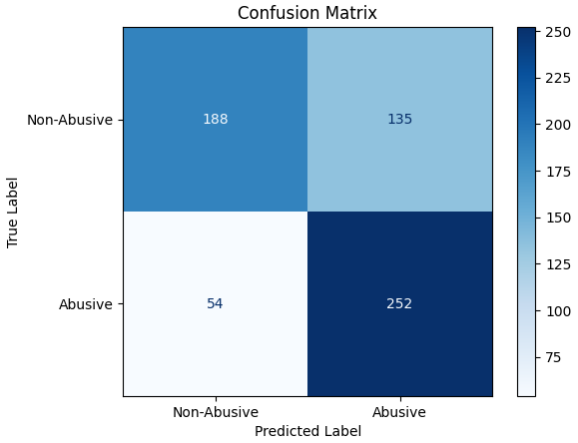


Figure 3: Confusion Matrix for Abusive and Non-Abusive Classification in Malayalam

and recall (0.74) for Non-Abusive content, resulting in more consistent performance.

As shown in Table 3, CUET_Agile achieved the highest mF1 score of 0.7883, followed closely by MSM_CUET (0.7873) and Incepto (0.7864). Our team, **byteSizedLLM**, achieved a Macro F1 score of 0.7820, ranking 6th. This demonstrates the effectiveness of our hybrid Attention BiLSTM-XLM-RoBERTa model in handling Tamil social media data. The close performance among the top teams suggests potential for further improvements through hyperparameter tuning and advanced data augmentation.

For the Malayalam dataset (Table 4), Habiba A, G Agila led with an mF1 score of 0.7571, followed by CUET_Agile (0.7234) and CUET_Novice (0.7083). Our team, **byteSizedLLM**, achieved a Macro F1 score of 0.6964, placing 6th. The lower performance for Malayalam reflects the linguistic diversity and complex code-mixed structures, highlighting the need for better class imbalance handling and transliteration strategies.

The model misclassified 135 abusive instances as non-abusive in Malayalam and 54 in Tamil, indicating better recall for abusive content in Tamil. Malayalam's non-abusive recall was 82%, effectively identifying non-abusive text but missing some abusive cases. Tamil's abusive recall was also 82%, reflecting a bias toward the majority class. Refer to Fig.2 for Tamil and Fig.3 for Malayalam.

The results show the effectiveness of our approach in detecting abusive comments in code-mixed multilingual text. The top team's strong performance highlights the potential of multilingual transformers, though improvements are needed for linguistic nuances and robustness to diverse code-mixed patterns

7 Limitations and Future Work

A key limitation was computational constraints, which restricted fine-tuning TamilXLM-RoBERTa and MalayalamXLM-RoBERTa models to a small subset of the AI4Bharat dataset. This hindered the model's ability to fully utilize the dataset's linguistic diversity and contextual richness.

For Tamil, the narrow performance gap among top teams indicates that transformer-based approaches have matured. However, for Malayalam, challenges like transliteration complexity and limited training data persist. Future work should address computational limits, explore pseudo-labeling and ensemble learning, and integrate external linguistic resources to improve performance.

8 Conclusion

This study demonstrated the potential of hybrid Attention BiLSTM-XLM-RoBERTa models for abusive comment detection in Tamil and Malayalam[5]. Despite computational constraints, our approach achieved competitive results, underscoring the effectiveness of integrating multilingual embeddings with sequential and attention mechanisms.

Future research should further refine these models by leveraging larger datasets, optimizing hyperparameters, and enhancing domain adaptation techniques to improve robustness and generalization.

[5] https://github.com/mdp0999/Abusive-Tamil-and-Malayalam-Text

References

S. Anbukkarasi and S. Varadhaganapathy. 2023. Deep learning-based hate speech detection in code-mixed tamil text. *IETE Journal of Research*, 69(11):7893–7898.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. AbhiPaw@ DravidianLangTech: Abusive comment detection in Tamil and Telugu using logistic regression. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 231–234, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Fazlourrahman Balouchzahi, Aparna B K, and H L Shashirekha. 2021. MUCS@DravidianLangTech-EACL2021:COOLI-code-mixing offensive language identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 323–329, Kyiv. Association for Computational Linguistics.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Bharathi Raja Chakravarthi, Mariappan Anandkumar, John P. McCrae, Bhavukam Premjith, K. P. Soman, and Thomas Mandl. 2020. Overview of the track on hasoc-offensive language identification-dravidiancodemix. In *Fire*.

Bharathi Raja Chakravarthi, Dhivya Chinnappa, Ruba Priyadharshini, Anand Kumar Madasamy, Sangeetha Sivanesan, Subalalitha Chinnaudayar Navaneethakrishnan, Sajeetha Thavareesan, Dhanalakshmi Vadivel, Rahul Ponnusamy, and Prasanna Kumar Kumaresan. 2021a. Developing successful shared tasks on offensive language identification for dravidian languages. *Preprint*, arXiv:2111.03375.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021b. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2022. Dravidiancodemix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, Elizabeth Sherly, John P. McCrae, Adeep Hande, Rahul Ponnusamy, Shubhanker Banerjee, and Charangan Vasantharajan. 2021c. Findings of the sentiment analysis of dravidian languages in code-mixed text. *Preprint*, arXiv:2111.09811.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Sobha Lalitha Devi. 2021. Anaphora resolution from social media text in indian languages (socanares-il)-overview. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '20, page 9–13, New York, NY, USA. Association for Computing Machinery.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Adeep Hande, Siddhanth U Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021a. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages. *Preprint*, arXiv:2108.03867.

Adeep Hande, Karthik Puranik, Konthala Yasaswini, Ruba Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. 2021b. Offensive language identification in low-resourced code-mixed dravidian languages using pseudo-labeling. *Preprint*, arXiv:2108.12177.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the*

Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and booktitle = Kumaresan, Prasanna Kumar". Findings of the shared task on Abusive Comment Detection in Tamil and Telugu.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021. Hatealert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 270–276, Kyiv. Association for Computational Linguistics.

Disne Sivalingam and Sajeetha Thavareesan. 2021. OffTamil@DravideanLangTech-EASL2021: Offensive language identification in Tamil text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 346–351, Kyiv. Association for Computational Linguistics.

Debapriya Tula, Prathyush Potluri, Shreyas Ms, Sumanth Doddapaneni, Pranjal Sahu, Rohan Sukumaran, and Parth Patwa. 2021. Bitions@DravidianLangTech-EACL2021: Ensemble of multilingual language models with pseudo labeling for offence detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 291–299, Kyiv. Association for Computational Linguistics.

Charangan Vasantharajan and Uthayasanker Thayasivam. 2021. Towards offensive language identification for tamil code-mixed youtube comments and posts. *SN Computer Science*, 3(1).

byteSizedLLM@DravidianLangTech 2025: Multimodal Misogyny Meme Detection in Low-Resource Dravidian Languages Using Transliteration-Aware XLM-RoBERTa, ResNet-50, and Attention-BiLSTM

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Abstract

Detecting misogyny in memes is challenging due to their multimodal nature, especially in low-resource languages like Tamil and Malayalam. This paper presents our work in the Misogyny Meme Detection task, utilizing both textual and visual features. We propose an Attention-Driven BiLSTM-XLM-RoBERTa-ResNet model, combining a transliteration-aware fine-tuned XLM-RoBERTa for text analysis and ResNet-50 for image feature extraction. Our model achieved Macro-F1 scores of 0.8805 for Malayalam and 0.8081 for Tamil, demonstrating competitive performance. However, challenges such as class imbalance and domain-specific image representation persist. Our findings highlight the need for better dataset curation, task-specific fine-tuning, and advanced fusion techniques to enhance multimodal hate speech detection in Dravidian languages.

1 Introduction

The proliferation of social media has transformed communication but has also led to the rise of harmful content, including misogynistic memes that combine text and visuals to convey discriminatory messages. Detecting such content is challenging due to its multimodal nature, implicit messaging, and linguistic diversity. Developing robust systems to identify and mitigate misogynistic memes is essential for fostering safer online spaces.

The Shared Task on Misogyny Meme Detection, part of DravidianLangTech@NAACL 2025[1], addresses this issue by focusing on memes in Tamil and Malayalam, two Dravidian languages with complex morphologies and distinct scripts. Participants are tasked with designing multimodal systems capable of analyzing both textual and visual components to classify memes as Misogynistic or Non-misogynistic. The challenges include handling transliterated text and capturing cultural nuances in linguistic expressions.

This task underscores the importance of multilingual and multimodal approaches in misogyny detection, particularly for Tamil and Malayalam, emphasizing culturally sensitive solutions in low-resource settings. Annotated social media datasets enable effective text and image processing, with a transliteration-aware fine-tuned XLM-RoBERTa-base handling textual content and ResNet-50 extracting visual features. This baseline serves as a foundation for exploring advanced architectures that integrate contextual information, with macro F1 score ensuring balanced evaluation across classes.

In this paper, we present our methodology, experimental setup, and results, demonstrating the effectiveness of our hybrid model in addressing the challenges of misogyny meme detection. We also discuss key challenges, such as transliteration, cultural nuances, and data sparsity, and propose directions for future research to enhance multilingual and multimodal misogyny detection.

2 Related Work

Advancements in multimodal image-text analysis have driven progress in hate speech detection, particularly with social media content. Early models like MOMENTA (Pramanick et al., 2021) and HateCLIPper (Kumar and Nandakumar, 2022) leveraged CLIP's vision-language encoders for cross-modal interactions, while newer methods refine alignment through textual inversion and image captioning.

MemeCLIP (Shah et al., 2024) directly utilizes CLIP's pre-trained encoders for meme processing, tackling data scarcity and class imbalance with Feature Adapters and a cosine classifier to enhance robustness.

[1] https://codalab.lisn.upsaclay.fr/competitions/20856

DravidianLangTech shared tasks highlight challenges in processing Tamil and Malayalam, especially with transliteration and code-mixing. Codewithzichao@DravidianLangTech-EACL2021 (Suryawanshi and Chakravarthi, 2021) employed XLM-RoBERTa and multilingual BERT for offensive language detection in Tamil, Malayalam, and Kannada, achieving strong F1 scores despite class imbalance (Li, 2021). Similarly, BPHC@DravidianLangTech-ACL2022 (V et al., 2022) focused on troll meme classification in Tamil-English code-mixed text, where MuRIL achieved a weighted F1 score of 0.74 (B et al., 2022).

Our work builds on these efforts, addressing misogyny meme detection in Tamil and Malayalam. We enhance CLIP's vision-language capabilities while tackling transliteration, code-mixing, and data sparsity, advancing multimodal analysis for Dravidian languages.

Label	Train	Dev	Test	Total
0	381	97	122	600
1	259	63	78	400
Total	640	160	200	1,000

Table 1: Statistics of the Malayalam Dataset for Misogynistic(1) and Non-Misogynistic(0) Classification

3 Dataset

The dataset for this task, provided as part of the Misogyny Meme Detection - DravidianLangTech@NAACL 2025 shared task(Ponnusamy et al., 2024), consists of multimodal memes in Malayalam and Tamil, annotated as misogynistic (1) or non-misogynistic (0). Each sample includes an image (JPG format) and transcribed text, with data split into train, development (dev), and test sets.

Table 1 and Table 2 present the data distribution for Malayalam and Tamil, respectively. This dataset benchmarks misogyny detection in low-resource languages, addressing challenges such as transliteration, code-mixing, and limited annotated data, fostering advancements in multilingual and multimodal learning.

4 Methodology

The proposed methodology utilizes a multimodal architecture to effectively handle textual and visual features for misogyny meme detection. The model combines the strengths of a transliteration-aware fine-tuned XLM-RoBERTa for text, ResNet-50 for image processing, and an attention-driven BiLSTM for multimodal feature fusion and classification.

Label	Train	Dev	Test	Total
0	851	210	267	1,328
1	285	74	89	448
Total	1,136	284	356	1,776

Table 2: Statistics of the Tamil Dataset for Misogynistic(1) and Non-Misogynistic(0) Classification

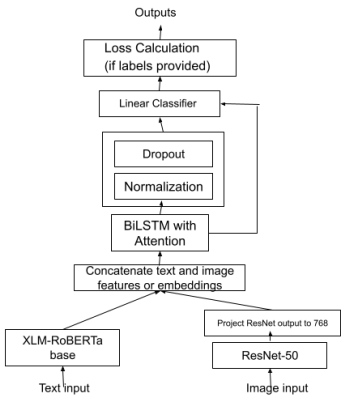


Figure 1: Architecture of the Attention-Driven BiLSTM-XLM-RoBERTa Classifier

4.1 XLM-RoBERTa Fine-Tuning with Transliteration Awareness

The XLM-RoBERTa Base model(Conneau et al., 2019) was fine-tuned on small portion of Tamil and Malayalam text from AI4Bharath(Kunchukuttan et al., 2020), achieving a perplexity of 4.9 for Tamil and 4.1 for Malayalam. To handle transliterated and mixed-script text, the IndicTrans tool(Bhat et al., 2015) was used to create three variations: native script, fully Romanized, and partially transliterated text. This preprocessing enhances text representation for diverse script inputs. The CLS token output provides a 768-dimensional embedding[2][3].

4.2 ResNet-50 for Image Feature Extraction

The visual component of the memes is handled using ResNet-50(He et al., 2016), a widely used convolutional neural network pre-trained on ImageNet. To align the visual features with the textual features, the fully connected (FC) layer of ResNet-50 is modified to project the extracted image features

[2] https://huggingface.co/bytesizedllm/TamilXLM_Roberta

[3] https://huggingface.co/bytesizedllm/MalayalamXLM_Roberta

into a 768-dimensional space. This modification ensures compatibility and seamless integration of textual and visual embeddings in the later stages of the model.

4.3 Attention-BiLSTM-XLM-RoBERTa-ResNet Classifier

We propose a hybrid Attention-Driven BiLSTM-XLM-RoBERTa-ResNet model for multimodal misogyny detection, inspired by our previous research (Kodali et al., 2025; Manukonda and Kodali, 2025, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b). This architecture integrates textual and visual features to capture both linguistic and image-based patterns.

The text input is processed using a fine-tuned XLM-RoBERTa, extracting contextual embeddings:

$$\mathbf{X}_t = \textbf{XLM-RoBERTa}(input\_ids, atten\_mask) \quad (1)$$

The image input is processed using ResNet-50 to extract deep visual features:

$$\mathbf{X}_i = \textbf{ResNet-50}(image\_features) \quad (2)$$

These features are concatenated (or element-wise added, averaged, attention-weighted, etc.) and passed through a BiLSTM layer (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) to capture sequential dependencies.

$$\mathbf{H}_t = [\mathbf{H}_{fwd,t}; \mathbf{H}_{bwd,t}] \quad (3)$$

An attention mechanism enhances key information:

$$\alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T} \exp(\mathbf{a}_t)} \quad (4)$$

$$\mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t \quad (5)$$

A fully connected layer classifies the output after layer normalization and dropout:

$$\textbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \quad (6)$$

The model is optimized using cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (7)$$

Figure 1 illustrates the model architecture, showcasing the integration of XLM-RoBERTa, ResNet-50, and BiLSTM with attention for enhanced multimodal classification.

5 Experimental Setup

This section describes the experimental setup for the proposed multimodal architecture, integrating XLM-RoBERTa for text, ResNet-50 for images, and BiLSTM with attention for feature fusion.

5.1 Text Processing

XLM-RoBERTa Base is fine-tuned on Tamil and Malayalam text to handle linguistic diversity in social media. IndicTrans preprocesses text into three formats: native script, fully Romanized, and partially transliterated (20–70%). Tokenization is performed with a max sequence length of 128, applying padding and truncation for batch uniformity.

5.2 Image Processing

ResNet-50, pre-trained on ImageNet, extracts visual features. The final layer is replaced with a projection layer to align 768-dimensional text embeddings. Images are resized to 224×224 pixels and normalized for consistency.

5.3 Multimodal Feature Fusion

Textual and visual embeddings are fused into a single tensor and processed via a BiLSTM, capturing cross-modal dependencies in a 512-dimensional space. An attention mechanism refines feature relevance before classification.

5.4 Training Configuration

The model is trained using AdamW with a learning rate of (learning rate 2×10^{-5} and a weight decay of 0.01, using a batch size of 16 for up to 5 epochs with early stopping. Cross-entropy loss is used for classification, and early stopping is applied based on the validation macro F1 score. Gradient clipping with a maximum norm of 1.0 ensures training stability.

6 Results and Discussion

Evaluation is based on macro F1, with accuracy and classification reports. The best macro F1 model is saved for testing. Our unique model setup for the shared task yielded notable results for both Tamil and Malayalam. For Malayalam, our second run[4] achieved a Macro F1 score of 0.8805, securing the highest score in the competition, surpassing the first-ranked team, CUET_Novice, which scored 0.8763. The best Macro F1 was obtained in our

[4] https://github.com/mdp0999/Misogyny-Meme-Detection/blob/main/test2_ml.ipynb

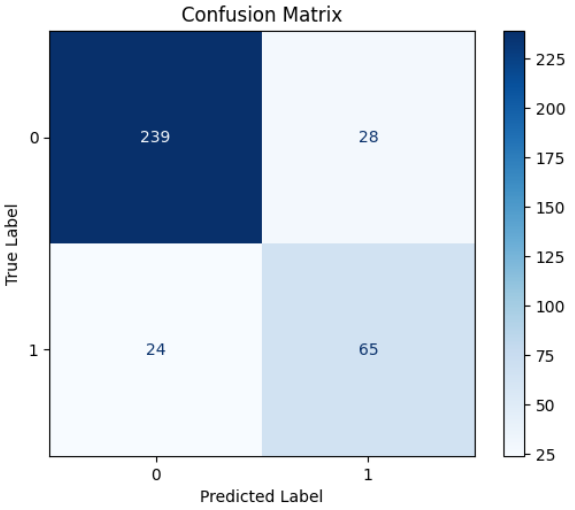


Figure 2: Confusion matrix for Task 1: Misogyny Meme Detection in Tamil.

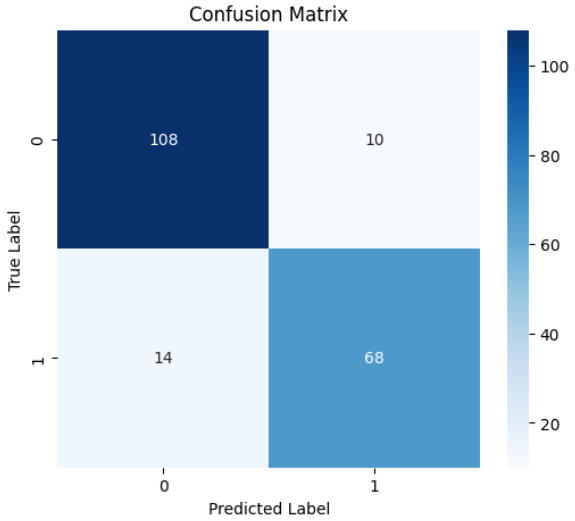


Figure 3: Confusion matrix for Task 2: Misogyny Meme Detection in Malayalam.

second run using a learning rate of 2×10^{-5}, while our first run, with a same learning rate and a customized transformer encoder-decoder architecture, resulted in a slightly lower score of 0.8391, as reflected in the task results.

In contrast, our Tamil model achieved a Macro F1 score of 0.8081, securing third place in the competition. The top score of 0.8368 for Tamil was achieved by team DLRG_RR. While our model performed well for the non-misogynistic class, its recall for the misogynistic class was lower, indicating challenges in capturing nuanced patterns associated with this class. These results suggest that class imbalance and limited training data may have hindered the model's ability to generalize effectively for Tamil.

The performance gap between Malayalam and Tamil suggests that class imbalance and script variations influenced misclassification. While the Malayalam model achieved high precision and recall across classes, the Tamil model struggled with lower recall for the misogynistic class, indicating difficulty in capturing nuanced linguistic patterns. The confusion matrices (Figures 2 and 3) highlight these challenges, emphasizing the need for better handling of class imbalance in Tamil and further refinement of feature extraction in both languages.

7 Limitations and Future Work

The primary limitation of our work was the restricted size of the training dataset due to computational constraints. This likely affected the model's ability to capture complex patterns, especially for the Tamil task. Additionally, the ResNet-50 architecture used for meme analysis was not fine-tuned for extracting features specific to misogyny memes, which may have limited its performance in visual understanding.

Future work will focus on addressing these limitations by training models on larger datasets to improve generalization and exploring meme-specific architectures for enhanced feature extraction. Furthermore, techniques to handle class imbalances more effectively will be incorporated to boost recall for minority classes. These advancements are expected to improve misogyny meme detection performance across multiple languages.

8 Conclusion

This study presents our approach to misogyny meme detection for Tamil and Malayalam languages, demonstrating strong performance for Malayalam with a top Macro F1 score of 0.8805 and competitive results for Tamil with a Macro F1 score of 0.8081. The findings emphasize the importance of addressing class imbalances, increasing data availability, and fine-tuning models for task-specific visual features. Despite its limitations, this work provides a robust foundation for future research and development in misogyny meme detection tasks. Our team, **byteSizedLLM**, remains committed to advancing solutions for such challenging multimodal tasks in low-resource languages.

References

Premjith B, Bharathi Raja Chakravarthi, Malliga Subramanian, Bharathi B, Soman Kp, Dhanalakshmi V, Sreelakshmi K, Arunaggiri Pandian, and Prasanna Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 254–260, Dublin, Ireland. Association for Computational Linguistics.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Zichao Li. 2021. Codewithzichao@DravidianLangTech-EACL2021: Exploring multilingual transformers for offensive language identification on code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 164–168, Kyiv. Association for Computational Linguistics.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. *Preprint*, arXiv:2409.14703.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132, Kyiv. Association for Computational Linguistics.

Achyuta V, Mithun Kumar S R, Aruna Malapati, and Lov Kumar. 2022. BPHC@DravidianLangTech-ACL2022-a comparative analysis of classical and pre-trained models for troll meme classification in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 151–157, Dublin, Ireland. Association for Computational Linguistics.

byteSizedLLM@DravidianLangTech 2025: Sentiment Analysis in Tamil Using Transliteration-Aware XLM-RoBERTa and Attention-BiLSTM

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Abstract

This study investigates sentiment analysis in code-mixed Tamil-English text using an Attention BiLSTM-XLM-RoBERTa model, combining multilingual embeddings with sequential context modeling to enhance classification performance. The model was fine-tuned using masked language modeling and trained with an attention-based BiLSTM classifier to capture sentiment patterns in transliterated and informal text. Despite computational constraints limiting pretraining, the approach achieved a Macro f1 of 0.5036 and ranked first in the competition. The model performed best on the Positive class, while Mixed Feelings and Unknown State showed lower recall due to class imbalance and ambiguity. Error analysis reveals challenges in handling non-standard transliterations, sentiment shifts, and informal language variations in social media text. These findings demonstrate the effectiveness of transformer-based multilingual embeddings and sequential modeling for sentiment classification in code-mixed text.

1 Introduction

Sentiment analysis involves identifying subjective opinions or emotions in text and has gained significant attention in both academia and industry. With the rise of social media, sentiment detection in Dravidian languages has become increasingly relevant, especially given the prevalence of code-mixing. Code-mixed texts, often written in non-native scripts, pose challenges for traditional monolingual sentiment analysis models due to complex linguistic variations and switching between languages.

The Shared Task on Sentiment Analysis in Tamil and Tulu at DravidianLangTech@NAACL 2025 focuses on message-level polarity classification of code-mixed Tamil-English and Tulu-English texts. Given a YouTube comment or post, the goal is to classify it as positive, negative, neutral, or mixed sentiment. The dataset, collected from social media, presents real-world challenges such as class imbalance and linguistic variability, necessitating robust NLP techniques for effective classification.

To address these challenges, we propose a transliteration-aware fine-tuning approach using XLM-RoBERTa, a state-of-the-art multilingual transformer model. The model is fine-tuned using Masked Language Modeling (MLM) on a subset of the AI4Bharath dataset(Kunchukuttan et al., 2020), incorporating original, fully transliterated, and partially transliterated text. This pretraining strategy equips the model to handle native scripts, Romanized text, and mixed-script data effectively.

Additionally, we integrate XLM-RoBERTa embeddings into a hybrid architecture with a attention-BiLSTM. The embeddings are projected and refined to capture complex contextual relationships in multilingual text. Dropout regularization and gradient clipping ensure stable training. Our approach achieves state-of-the-art performance, demonstrating the effectiveness of transliteration-aware pretraining and hybrid architectures in handling sentiment classification for code-mixed Dravidian languages.

This study analyzes data preprocessing, MLM training, and classifier design, introducing innovations that improve detection accuracy and scalability. The proposed framework enhances Sentiment Analysis in Tamil, providing insights into model performance and deployment challenges.

2 Related Work

Sentiment Analysis on social media has progressed significantly, with growing attention to low-resource languages like Tamil. Chakravarthi et al. (2021) and B et al. (2022) organized shared tasks to promote Sentiment Analysis in code-mixed Dravidian languages, laying a foundation for tackling linguistic diversity in Sentiment Analysis.

Several approaches have explored Tamil-English code-mixed data. S R et al. (2022) addressed data imbalance using kernel-based learning and advanced feature selection techniques, while Shanmugavadivel et al. (2022) employed hybrid deep learning models, combining CNN and BiLSTM architectures, achieving strong results for mixed-language datasets. Preprocessing steps, including emoji and punctuation removal, and TF-IDF-based feature extraction, were crucial to their success.

Sentiment Analysis in Tamil has been explored through various shared tasks, such as DravidianLangTech@RANLP 2023 (Priyadharshini et al., 2023; Hegde et al., 2023a) and DravidianLangTech@EACL 2024 (Sambath Kumar et al., 2024). In 2023, XLM-RoBERTa with adversarial and ensemble training demonstrated the effectiveness of transformers for Tamil-English code-mixed text (Luo and Wang, 2023). The task also addressed abusive language detection in Tamil, Telugu, and Tamil-English code-mixed texts using approaches like LinearSVC with n-grams and Transfer Learning models with BERT variants, emphasizing the ongoing challenges in handling abusive content effectively (Hegde et al., 2023b).

In 2024, B et al. (2024) implemented SVM and an ensemble of ML classifiers—Support Vector Model (SVM), Random Forest (RF), and k Nearest Neighbors (kNN)—for Tamil-English code-mixed sentiment analysis. They used GridSearch for hyperparameter tuning, achieving a top macro F1 score and securing a top rank in the shared task.

Despite these advancements, Sentiment Analysis in Tamil remains an open challenge, requiring further improvements in methods and performance.

3 Dataset

The dataset for Sentiment Analysis in Tamil task consists of code-mixed Tamil-English comments and posts collected from social media platforms. Each instance is annotated with one of four sentiment labels: Positive(0), Negative(1), Mixed Feelings(2), and Unknown State(3). The dataset presents class imbalance, reflecting real-world sentiment distribution in online discourse(Chakravarthi et al., 2020).

The data is divided into training, validation, and test sets, ensuring a robust benchmark for sentiment classification. This is a message-level polarity classification task, and Table 1 summarizes the dataset distribution.

Label	Train	Val	Test	Total
0	18145	2272	1983	22400
1	4151	480	458	5089
2	3662	472	425	4559
3	5164	619	593	6376
Total	31122	3843	3459	38424

Table 1: Dataset distribution across sentiment labels in Train, Validation, and Test splits.

This dataset serves as a benchmark for exploring sentiment expression in code-mixed Tamil-English text, addressing challenges such as transliteration, informal language, and code-switching in social media discourse.

4 Models

This section presents the models used in our experiments. Fine-tuned XLM-RoBERTa with Masked Language Modeling (MLM) enhances processing of Tamil-English code-mixed text. An attention-driven BiLSTM further refines embeddings, improving contextual understanding and sequence modeling.

4.1 Fine-Tuning XLM-RoBERTa with MLM

XLM-RoBERTa, a multilingual transformer model based on RoBERTa, is trained on a large-scale Common Crawl corpus spanning 94 languages (Conneau et al., 2019). It employs dynamic masking and optimized pretraining, enabling it to capture complex linguistic patterns across languages.

To enhance its ability to process transliterated and code-switched Tamil-English text, we fine-tuned the base XLM-RoBERTa model using Masked Language Modeling (MLM). This pretraining strategy involves masking random tokens and training the model to predict them, allowing it to learn robust contextual embeddings tailored to bilingual text.

The MLM training dataset was constructed from monolingual Tamil social media text, fully transliterated text in Roman script, and partially transliterated text with 20–70% of words transliterated. This approach enabled the model to recognize native script, Romanized text, and mixed-script data, crucial for processing real-world Tamil-English social media content.

The fine-tuned XLM-RoBERTa model (TamilXLM_Roberta[1]) serves as the embedding

[1] https://huggingface.co/bytesizedllm/TamilXLM_

backbone for sentiment classification, enhancing its ability to handle linguistic and orthographic variability in code-mixed datasets.

4.2 Attention BiLSTM-XLM-RoBERTa Model

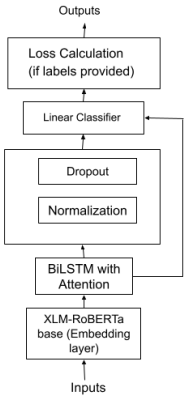


Figure 1: Architecture of the BiLSTM-XLM-RoBERTa Classifier Model.

This study introduces a hybrid Attention BiLSTM-XLM-RoBERTa model for multi-label classification, integrating fine-tuned XLM-RoBERTa embeddings with a BiLSTM and attention mechanism (Liu and Guo, 2019; Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Kodali et al., 2025; Manukonda and Kodali, 2025, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b). As shown in Figure 1, the model captures contextual dependencies using BiLSTM and assigns dynamic importance to hidden states via attention.

XLM-RoBERTa generates contextual embeddings, which are processed by BiLSTM to extract forward and backward hidden states. An attention mechanism computes weight distributions to refine the representation:

$$\mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t, \quad \alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T} \exp(\mathbf{a}_t)} \tag{1}$$

Residual components such as layer normalization and dropout are applied to the attention-weighted representation to stabilize training and reduce overfitting:

Roberta

$$\mathbf{H}_{dropout} = Dropout(LayerNorm(\mathbf{H}_{attended})) \tag{2}$$

Finally, a classification layer outputs logits:

$$\textbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \tag{3}$$

The model is trained using cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{4}$$

This architecture effectively combines XLM-RoBERTa embeddings, BiLSTM, and attention to enhance multi-label classification in code-mixed text.

5 Experiment Setup

The experiments evaluate the integration of attention-based BiLSTM with fine-tuned XLM-RoBERTa embeddings for sentiment analysis in code-mixed Tamil-English text. XLM-RoBERTa was fine-tuned using Masked Language Modeling (MLM) with a 15% masking probability, a batch size of 16, and a learning rate of 5×10^{-5}. The model was trained for up to ten epochs with early stopping based on validation perplexity.

For classification, the fine-tuned embeddings were processed through a BiLSTM model with two LSTM layers (hidden size 512) and an attention mechanism to enhance contextual representation. A dropout probability of 0.3 was applied for generalization. The model was trained using AdamW with a learning rate of 2.5×10^{-5} and weight decay of 0.01, running for six epochs with early stopping based on validation loss and macro F1-score.

This setup demonstrates the effectiveness of combining XLM-RoBERTa embeddings, BiLSTM, and attention mechanisms for sentiment classification in Tamil-English text, addressing challenges such as transliteration, informal language, and linguistic variability in social media data.

6 Results and Discussion

XLM-RoBERTa achieved a perplexity of 4.9 for Tamil bilingual text, indicating its effectiveness in modeling code-mixed language representations.

The performance of the Attention BiLSTM-XLM-RoBERTa model was evaluated on the code-mixed Tamil-English sentiment analysis task[2]. The

[2] https://github.com/mdp0999/Sentiment-Analysis-in-Tamil

Label	Precision	Recall	F1-Score	Support
Mixed Feelings	0.27	0.24	0.26	425
Negative	0.53	0.46	0.49	458
Positive	0.76	0.84	0.79	1983
Unknown State	0.51	0.41	0.45	593
Accuracy	-	-	0.64	3459
Macro Avg	0.52	0.49	0.50	3459
Weighted Avg	0.62	0.64	0.63	3459

Table 2: Classification Report on the Test Set for Sentiment Analysis in Code-Mixed Tamil-English Text

classification report in Table 2 shows an overall accuracy of 64 percent, with a macro F1-score of 0.50 and a weighted F1-score of 0.63.

The model performed best on the Positive class, achieving an F1-score of 0.79. This can be attributed to the higher representation of Positive instances in the dataset, allowing the model to learn its distinguishing features more effectively. In contrast, the Mixed Feelings and Unknown State categories had lower F1-scores of 0.26 and 0.45, respectively, suggesting difficulty in distinguishing ambiguous sentiment. The Negative class obtained a moderate F1-score of 0.49, reflecting challenges in identifying negative sentiment, which often overlaps with neutral or mixed sentiments.

Several misclassifications stemmed from class imbalance, ambiguous sentiment expressions, and code-switching complexity. Mixed Feelings and Unknown State were often misclassified as Positive or Negative due to overlapping linguistic cues, especially in subtle or sarcastic expressions. Errors also arose from transliteration inconsistencies, as Tamil-English text lacks standardized spelling. Variations in transliteration, spelling errors, and informal language led to confusion, affecting sentiment assignment. The model also struggled with sentiment shifts in longer sentences, resulting in incorrect predictions when sentiment changed mid-sentence.

Team	Score	Rank
byteSizedLLM	**0.5036**	**1**
ET2025	0.4986	2
Hermes	0.4957	3
JustATalentedTeam	0.4919	4
Lemlem	0.4709	5

Table 3: Performance ranking of different teams based on their submitted runs.

7 Limitations and Future Work

This study was limited by computational constraints, restricting XLM-RoBERTa pretraining and its generalization to diverse Tamil-English code-mixed patterns. Class imbalance, particularly in the Mixed Feelings and Unknown State categories, led to biased classification. Additionally, low transliteration accuracy introduced inconsistencies in text representation, affecting sentiment detection. Addressing these challenges through data augmentation techniques such as back-translation and oversampling could improve recall for underrepresented classes.

Future work will focus on developing more accurate transliteration models for code-mixed text, improving representation consistency. Expanding pretraining on larger datasets and overcoming computational limitations could enhance model performance. Additionally, integrating multimodal data and exploring domain adaptation techniques may improve robustness in handling informal and noisy social media text.

8 Conclusion

This study presented an Attention BiLSTM-XLM-RoBERTa model for sentiment analysis in code-mixed Tamil-English text, effectively capturing sentiment cues by leveraging multilingual embeddings and sequential modeling. The model achieved competitive performance, but challenges such as class imbalance, ambiguous sentiment transitions, and low transliteration accuracy affected classification of underrepresented categories. Error analysis highlighted the need for improved handling of informal and transliterated text. Future enhancements, including better pretraining, data augmentation, and robust transliteration models, can further refine sentiment detection in code-mixed social media text.

References

Prathvi B, Manavi K, Subrahmanyapoojary K, Asha Hegde, Kavya G, and Hosahalli Shashirekha. 2024. MUCS@DravidianLangTech-2024: A grid search approach to explore sentiment analysis in code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 257–261, St. Julian's, Malta. Association for Computational Linguistics.

Premjith B, Bharathi Raja Chakravarthi, Malliga Subramanian, Bharathi B, Soman Kp, Dhanalakshmi V, Sreelakshmi K, Arunaggiri Pandian, and Prasanna Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 254–260, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2021. Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '20, page 21–24, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, Lavanya S K, Thenmozhi D., Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023a. Findings of the shared task on sentiment analysis in Tamil and Tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Asha Hegde, Kavya G, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2023b. MUCS@DravidianLangTech2023: Leveraging learning models to identify abusive comments in code-mixed Dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 266–274, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Zhipeng Luo and Jiahui Wang. 2023. DeepBlueAI@DravidianLangTech-RANLP 2023. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 171–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and

streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mithun Kumar S R, Lov Kumar, and Aruna Malapati. 2022. Sentiment analysis on code-switched Dravidian languages with kernel based extreme learning machines. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 184–190, Dublin, Ireland. Association for Computational Linguistics.

Lavanya Sambath Kumar, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 2024. Overview of second shared task on sentiment analysis in code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 62–70, St. Julian's, Malta. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Comput. Speech Lang.*, 76(C).

SSNCSE@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages

Sreeja K, Bharathi B
Department of Computer Science and Engineering
Sri Sivasubramania Nadar College of Engineering
sreeja2350625@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Hate speech detection is a serious challenge due to the different digital media communication, particularly in low-resource languages. This research focuses on the problem of multimodal hate speech detection by incorporating both textual and audio modalities. In the context of social media platforms, hate speech is conveyed not only through text but also through audio, which may further amplify harmful content. In order to manage the issue, we provide a multiclass classification model that influences both text and audio features to detect and categorize hate speech in low-resource languages. The model uses machine learning models for text analysis and audio processing, allowing it to efficiently capture the complex relationships between the two modalities. The class weight mechanism involves avoiding overfitting. The prediction has been finalized using the majority fusion technique. Performance is measured using a macro average F1 score metric. Three languages—Tamil, Malayalam, and Telugu—have the optimal F1 scores, which are 0.59, 0.52, and 0.33.

1 Introduction

In the digital era, the analysis of multimodal social media data aligns your insights with very different types of diverse data appearing on social networks, including text, audio, and video. However, with the advent of social networks, platforms such as YouTube, Facebook, and Twitter not only aided in information sharing and networking, but also became a place where people were targeted, defamed, and marginalized based on their religion, sex, political, and personal defamation. Social networks have become increasingly integrated in this digital age; it has changed the perception of networking and socializing.

Not only humans, but chatbots can also corrupted by hate speech content. After learning foul language from user interactions, Microsoft's chatbot "Tay"(Neff and Nagy, 2016), which was designed to engage people in lighthearted and informal discussion, began using it. The hate content was too obvious for the chatbot to identify and avoid. This serves as an illustration of how important it is to identify hate speech in tweets and social networks for applications such as sentiment analysis, chatbot development, content recommendation, etc. An efficient identification guarantees a safer, more moral AI system and a blocking mechanism against the spreading of dangerous content.

Hate speech analysis models trained for such contexts must reflect features of all modalities concerned. In our case, the task is to classify multimodal (text and audio) data in Tamil, Malayalam, and Telugu into five separate hate classes: gender, political, religious, personal defamation, and non-hate.

The rest of the paper is organized as follows: Section 2 analyzes the related works done in the previous research, and Section 3 discusses the hate speech corpus in the current work. Section 4 contains a detailed discussion of the proposed models used in the current work. Section 5 explains the experimental results. Section 6 discusses the limitations. In Section 7, concludes the paper.

2 Related works

Detecting hate speech is the most effective way to make any environment safe, inclusive, and respectful, both online and offline. This will protect individuals from emotional distress, psychological suffering, and the risky transition from hostility to physical harm. The rate of division is decreased along with social integration and tolerance in communities when hate speech is recognized and suppressed. However, hate speech detection

in low-resource languages is challenging due to limited linguistic resources, the complexity and dynamic systems of cultures, and technological gaps. All of these challenges need strong documented work in collecting data, culturally sensitive models, and tailored approaches for fairness and effectiveness. (Lal G et al., 2025)provides an overview of the shared task on Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL). The paper explores multiclass hate speech detection in Dravidian languages. Detecting hate content in social media comments is not a novel concept for the English language(Kumar and Singh, 2022)(Jemima et al., 2022). Several systems have also been developed for languages other than English, such as Hindi, German (Rajalakshmi et al., 2022),(Rajalakshmi and Reddy, 2020). However, limited research focuses on identifying offensive content in low-resource Dravidian languages such as Tamil, Malayalam, and Kannada(Roy et al., 2022). The study proposes a method for identifying hate speech in low-resource languages in Tamil, Malayalam, and Telugu. The proposed model expands the task into multiclass classification, with the intent of detecting hate speech in various categories to refine the classification and enhance the detection. The switch from binary to multiclass classification identifies the potential of hate speech across different contexts and modalities. To improve the approach of (Boishakhi et al., 2021), and (Premjith et al., 2024b) which initially employed binary classification, we extend the method to handle multiclass classification. Moreover, the Binary class distinguishes only hate and non-hate. In contrast, multiclass classification categorizes the content in its target or intent, providing a deeper understanding of why it is considered hateful. This approach uses multiclass categorization for the detection and classification to prioritize and identify hate speech types. In multiclass classification the classes are imbalanced, to overcome this(Sreelakshmi et al., 2024) uses the class weight mechanism by assigning more weights to minority classes and the model pays more attention to them. Multimodal classification for abusive comment detection was discussed in (Anierudh et al., 2024).

3 Dataset Description

The task aims to develop a model for detecting Hate speech in low-resource languages namely Tamil, Malayalam, and Telugu. The dataset is sourced from the Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) provided by DravidianLangTech@NAACL 2025(Premjith et al., 2024b),(Premjith et al., 2024a). The task comprises three subtasks, and each subtasks contains two modalities data like Text and Audio. Each Audio data has a corresponding Transcript in Text data. The subtasks are Multimodal Hate Speech Detection in low-resource languages namely Tamil, Malayalam, and Telugu. Each language contains 514, 863, and 556 training samples for Tamil, Malayalam, and Telugu, as well as 50 test samples for each. This is a multi-class classification task, the classes are Gender (G), Political (P), Religious (R), Personal Defamation(C), and Non-Hate(N).

Category	Tamil	Malayalam	Telugu
Non-Hate (N)	287	406	198
Personal Defamation (C)	65	186	122
Gender (G)	68	82	106
Political (P)	33	118	58
Religious (R)	61	91	72
Total	514	863	556

Table 1: Training dataset distribution for Tamil, Malayalam, and Telugu.

4 Proposed Methodology

This study proposes a systematic methodology for classifying multimodal data encompassing four low-resource languages. The entire study is divided into five stages: data preprocessing, data balancing, feature extraction, classifier modeling, and majority fusion mechanism for predictions. Each stage has been thoughtfully designed to counter various problems posed by multimodal and low-resource language data.

Data preprocessing is the most important step in preparing the input multimodal data. The data have different modalities (text and audio) that may be inconsistent or noisy. The preprocessing pipeline consists of the following steps: cleaning the data by removing noisy information, and normalizing modalities to ensure consistency. The textual data undergo techniques such as tokenization and processing by language-specific methods. This step ensures a clean, structured, and aligned dataset ready for further processing.

Class weights mechanism has been utilized to balance the classes. Class weight was guaranteed

to provide more importance for minority classes during training. The class imbalance prevents the classifiers from biasing towards majority classes by improving performance.

Feature extraction serves as the most important for training the data. Textual data is vectorized using TF-IDF vectorizer and CountVectorizer. For audio data, MFCC and Log-mel spectrograms are employed. This step guarantees that diverse modalities are successfully transformed into feature-modeling vectors that can feed into machine-learning models.

The research proposes a classifier training method. Classifier models are used to perform the multiclass classification task: Support Vector Machines(SVM), Random Forests(RF), Multi-layer Perceptron(MLP) classifier, and Logistic Regression(LR). SVM uses kernel functions to handle linear as well as non-linear relationships effectively. Random Forest is another class of techniques that exploits an ensemble-based approach, which is very robust in capturing feature interactions. The MLP classifier is a feedforward artificial neural network with input, hidden, and output layers, among other layers. It uses backpropagation for training and applies activation to capture non-linear relationships in the data. It can handle structured data effectively due to its wide versatility. Logistic Regression is a highly popular classification task, it is a simple but effective linear model for providing great interpretability and strong baseline performance. Each model is trained successfully.

Multimodal data aggregation is performed using a majority fusion mechanism to combine predictions across modalities and models. A majority voting method is used for the first time to merge the predictions of three classifiers from each modality. The final output for each modality is taken as the label predicted by the majority of the classifiers for an instance of each modality. The results of all modalities are fused again using another majority voting mechanism to produce the overall model prediction. This two-level fusion mechanism ensures that all artifacts from all modalities and classifiers are substantially fused to obtain a robust and accurate prediction system.

In summary, the proposed methodology constitutes a complete workflow for classifying multimodal data for low-resource language speakers. The issues of sparsity, imbalance, and multimodal integration are directly addressed by including preprocessing, class balancing, modality-specific feature extraction, classifier selection, and hierarchical majority fusion. In other words, the final majority fusion takes place based on modality-wise predictions, which helps the model to draw on the diverse strengths of the classifiers and modalities.

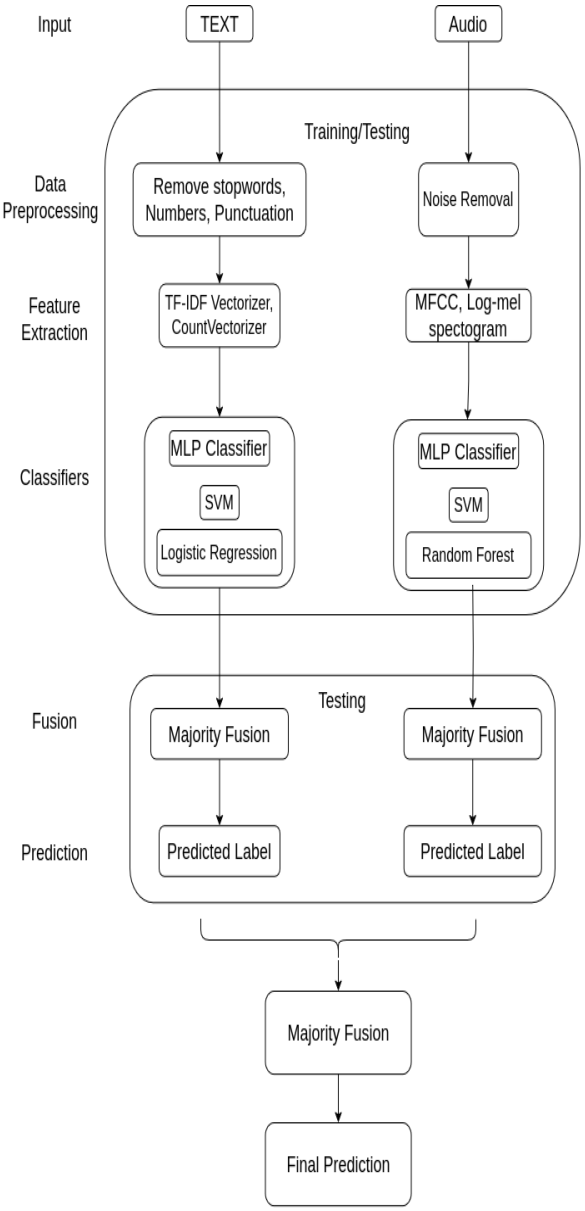


Figure 1: Architecture Diagram of the proposed work

5 Experimental Results

The performance of the Multimodal Hate Speech Detection model was evaluated with a macro F1-score. Text is trained with SVM, MLP classifier, and Logistic Regression model for feature extraction TF-IDF and Count vectorizer are used, and audio is trained with SVM, MLP classifier, and Random forest model for feature extraction MFCC and Log-mel Spectrogram are used. Using Majority Fusion technique Text and audio is fused independently. Finally, text and audio are both fused in

the Majority fusion mechanism.

Increasing the weight of a particular model in an ensemble learning system can result in huge improvements in performance, Table 2 shows the improved result, whereas Table 3 shows the results without any weight adjustment. In the weighted method described in Table 2, better-performing models will have additional influence on the final prediction, thus giving rise to better results. On the contrary, in Table 3, the models are treated uniformly; this may affect the overall accuracy downwards as well as result in less influence from more accurate models. The source code for the proposed approach and found here [1].

Metric	Tamil	Malayalam	Telugu
Accuracy	0.50	0.42	0.38
F1-Score	0.47	0.34	0.34
Precision	0.47	0.39	0.36
Recall	0.50	0.42	0.38

Table 2: Performance analysis of multimodal hate speech detection across languages without class weights

Metrics	Tamil	Malayalam	Telugu
Accuracy	0.60	0.56	0.38
F1-Score	0.59	0.52	0.33
Precision	0.63	0.57	0.33
Recall	0.60	0.56	0.38

Table 3: Performance analysis of multimodal hate speech detection across languages using class weights

6 Limitations

The current work caught several hurdles, including:

- The paper acknowledges the cruciality of obtaining sufficient and representative data for detecting hate speech in low-resource languages. The model suggests generalizability and strength may be impacted by this constraint.
- The issue of imbalance in hate speech datasets is reduced by using the application of an imbalance class weights technique, biases in the predictions will still exist to some extent, especially when it comes to the minority classes.
- The model becomes more sophisticated as text and audio modalities are added. The two-level fusion technique proposed in the research still requires additional testing before it can be used in real-world scenarios.

[1] https://github.com/SreejaKumaravel/Multimodal-Hate-Speech-Detection

7 Conclusions

The study concludes with a demonstration of the effectiveness of the majority fusion and class weighing in machine learning models for multimodal hate speech identification. In multiclass classification tasks, weighted-class models are preferred because they satisfy underrepresented classes and become sensitive enough to these class instances. With robust fusion methods capable of combining different model outputs, it is likely to obtain the optimal F1 score, which is one of the most important metrics in evaluating classification performance on imbalanced datasets. The experimental results show the promise of this method in dealing with the challenge of multimodal data and unbalanced class distribution and may lead to future advances in hate speech detection systems.

References

S Anierudh, R Abhishek, Ashwin Sundar, Amrit Krishnan, and B Bharathi. 2024. Wit hub@ dravidianlangtech-2024: Multimodal social media data analysis in dravidian languages using machine learning models. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 229–233.

Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md Golam Rabiul Alam. 2021. Multi-modal hate speech detection using machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4496–4499. IEEE.

P. Preethy Jemima, Bishop Raj Majumder, Bibek Kumar Ghosh, and Farazul Hoda. 2022. Hate speech detection using machine learning. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pages 1274–1277.

Gunjan Kumar and Jyoti Prakash Singh. 2022. Hate speech and offensive content identification in english and indo-aryan languages using machine learning models. In *FIRE (Working Notes)*, pages 542–551.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Gina Neff and Peter Nagy. 2016. Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 10:4915–4931.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

R Rajalakshmi and Yashwant Reddy. 2020. An enhanced ensemble classifier for hate and offensive content identification. *Journal of E-Technology Volume*, 11(2):71.

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022. DLRG@DravidianLangTech-ACL2022: Abusive comment detection in Tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213, Dublin, Ireland. Association for Computational Linguistics.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech Language*, 75:101386.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

YenCS@DravidianLangTech 2025: Integrating Hybrid Architectures for Fake News Detection in Low-Resource Dravidian Languages

Anusha M D
Department of Computer Science,
Yenepoya Institute of Arts,
Science, Commerce and Management,
Yenepoya (Deemed to be University),
Balmata, Mangalore
anugowda251@gmail.com

Parameshwar R Hegde
Department of Computer Science,
Yenepoya Institute of Arts,
Science, Commerce and Management,
Yenepoya (Deemed to be University),
Balmata, Mangalore
param1000@yahoo.com

Abstract

Detecting fake news in under-resourced Dravidian languages is a rigorous task due to the scarcity of annotated datasets and the intricate nature of code-mixed text. This study tackles these issues by employing advanced machine learning techniques for two key classification tasks, the first task involves binary classification achieving a macro-average F1-score of 0.792 using a hybrid fusion model that integrates Bidirectional Recurrent Neural Network (Bi-RNN) and Long Short-Term Memory (LSTM)-Recurrent Neural Network (RNN) with weighted averaging. The second task focuses on fine-grained classification, categorizing news where an LSTM-GRU hybrid model attained a macro-average F1-score of 0.26. These findings highlight the effectiveness of hybrid models in improving fake news detection for under-resourced languages. Additionally, this study provides a foundational framework that can be adapted to address similar challenges in other under-resourced languages, emphasizing the need for further research in this area.

Keywords: Dravidian Languages, Fake News Detection, Hybrid Models, Multi-Class Classification

1 Introduction

Fake news consists of misleading or false information that imitates the structure and style of authentic news (Devika et al., 2024). Its spread can cause substantial societal misperceptions, sometimes resulting in severe consequences. Hence, distinguishing genuine news from fake news is essential.

If news is inaccurate, it can mislead individuals and contribute to the dissemination of false information (Subramanian et al., 2025). In some cases, fake news is deliberately used to generate rumors or damage the reputation of political figures (Subramanian et al., 2023). To address this challenge, a system has been proposed for detecting fake news. However, given the vast volume of data available on the internet and social media, manually verifying the authenticity of news content remains a significant challenge(Yigezu et al., 2023).

This widespread phenomenon spreads rapidly, affecting a vast number of people on a daily basis. The far-reaching influence of fake news presents substantial risks to national security, economic stability, and public welfare. Regrettably, many individuals remain unaware of the profound consequences fake news can have on crucial issues and often lack the necessary skills to identify and mitigate such challenges(Yigezu et al., 2023). The study was conducted by the organizers of a shared task, which involves two distinct tasks: classifying social media text as either original or fake, and identifying multiple labels in Malayalam news (Subramanian et al., 2024). This research aims to investigate the effectiveness of different machine learning, deep learning, and hybrid models in tackling critical challenges in text classification tasks. Through the use of advanced techniques and optimization of model architectures, the study seeks to contribute to the development of robust solutions for processing complex datasets, with a particular focus on under-resourced and code-mixed languages.

2 Literature Review

In terms of feature extraction, contextual understanding, and enhancing classification accuracy, RNNs and LSTMs have demonstrated remarkable efficacy Waqas and Humphries (2024). With an emphasis on using deep learning frameworks to process complex textual data, this section gives a summary of recent developments in binary and multi-class classification tasks.

Numerous studies have emphasized the potential of deep learning methods for addressing diverse classification tasks. Yigezu et al. (2024) employed an RNN-LSTM model, with hyperparameters opti-

mized via grid search. The model demonstrated notable effectiveness in binary classification, achieving an accuracy of 0.82. However, its performance on multi-class tasks was compromised due to the issue of imbalanced data, resulting in a lower score of 0.32. Similarly,Chauhan and Palivela (2021) applied an LSTM-based approach for fake news detection, utilizing GloVe word embeddings to represent text as vectors, tokenization for feature extraction, and N-grams to enhance feature representation. When compared to Alghamdi et al. (2022) fake news detection methods, their model achieved an outstanding accuracy of 99.88%, highlighting the strength of LSTM networks in processing complex textual data and distinguishing between false and genuine news.

Convolutional neural networks (CNN) and recurrent neural networks with long short-term memory (RNN-LSTM) were combined in Goonathilake and Kumara (2020) to create a hybrid model for text classification. Convolution and max-pooling were used by the CNN to extract features, and the RNN-LSTM to record long-term dependencies. Overfitting was lessened by dropout regularization and dense layers, and the Adam optimizer with binary cross-entropy loss attained 92% accuracy.

3 Methodology

The methodology describes the structured process followed to identify fake news in Dravidian languages. This process encompasses text preprocessing, tokenization, and padding, which are crucial steps for preparing the data for analysis and efficient model training.

3.1 Dataset

The dataset employed in this research originates from the Shared Task on Fake News Detection in Dravidian Languages, organized at Dravidian-LangTech@NAACL 2025.

Table 1 and Table 2 provide the class-wise distribution of the dataset both the task respectively. This dataset is crucial for advancing fake news detection in Dravidian languages, a less explored area in computational linguistics.

Classes	Train	Test	Dev
Original	1658	512	409
Fake	1599	507	406
Total	**3257**	**1019**	**815**

Table 1: Class-wise Distribution of Dataset for Task A

Classes	Train Set	Test Set
Half True	145	24
False	1,251	149
Partly False	44	14
Mostly False	242	63
Total	**1,682**	**250**

Table 2: Class-wise Distribution of Dataset for Task B

3.2 Pre-processing

The pre-processing pipeline involves tokenization using the Keras Tokenizer[1] after the data has been cleaned up by eliminating stopwords, mentions, punctuation, and numbers. In order to maintain consistent input dimensions, sequences are then padded to 100 words. Lastly, for multi-class classification, labels are encoded using one-hot encoding method.

3.3 Feature Extraction

Semantic representations of words in the dataset are derived using FastText embeddings, as detailed in FastText. FastText, an extension of Word2Vec (Church, 2017), represents words as bags of character n-grams, enabling it to generate meaningful embeddings even for out-of-vocabulary words. Each word in the Malayalam text is mapped to a dense vector using pre-trained FastText embeddings. This approach enhances the performance of subsequent machine learning and deep learning models by improving their understanding of language patterns in code-mixed text(Umer et al., 2023).

3.4 Model Building

1. Task A: Binary Classification

   In this task, a hybrid model approach combined with an ensemble strategy using weighted averaging, known as the fusion model (Alyahyan, 2025), is employed to classify social media posts, particularly YouTube comments, into fake or original categories.

   - Bi-RNN: uses tokenized and padded sequences, with a pre-trained embedding layer and a Bi-RNN layer (128 units) to capture contextual dependencies in both forward and backward directions. The output is passed through a dense layer with ReLU activation before the final classification layer. This structure allows the model to effectively learn from

[1]https://keras.io/keras_hub/api/tokenizers/tokenizer/

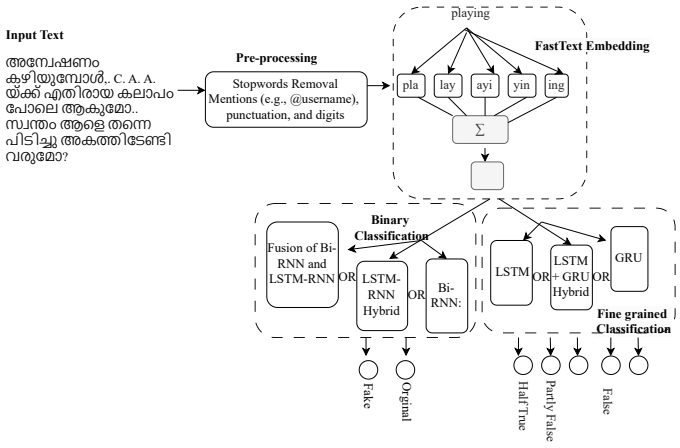


Figure 1: Proposed Methodology for Binary and Fine-Grained Classification

sequential data with complex dependencies(Yang et al., 2022).

- LSTM-RNN Hybrid: combines an embedding layer, an LSTM layer (128 units), and a Simple RNN layer (64 units) to extract sequential features(Telmem et al., 2024). The output is processed by a dense layer for classification.
- Fusion of Bi-RNN and LSTM-RNN (Ensemble Method): The predictions from the Bi-RNN and LSTM-RNN models are integrated through weighted averaging, utilizing the distinct advantages of each model to enhance performance. (Telmem et al., 2024). Figure 2 illustrates the fusion model.

  Experiments with ensembles of DNN, LSTM, RNN, and GRU models were also conducted. The predictions from each model were weighted and aggregated, enhancing robustness and performance in classification tasks.

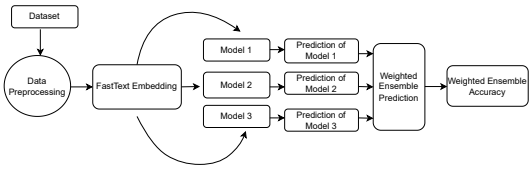


Figure 2: Architecture of the Fusion Model

2. Task B: Fine-grained classification

   The methodology focuses on detecting and classifying fake news in Malayalam-language news articles. This is achieved by employing a range of advanced model architectures to categorize the articles into five predefined categories.

   - GRU: employs an embedding layer initialized with a pre-trained embedding matrix, followed by a Conv1D layer (128 filters) with ReLU activation, and a MaxPooling1D layer for downsampling. The model includes a GRU layer (128 units) to capture sequential dependencies, followed by a Dropout layer (0.2) for regularization(Xu, 2024). The output layer uses softmax activation to handle the multi-class classification task with 5 output classes.
   - LSTM + GRU Hybrid: begins with an embedding layer, followed by an LSTM layer (128 units) with return_sequences=True to capture long-term dependencies. A GRU layer (64 units) processes the LSTM output to extract further sequential features Mousa et al. (2024). After applying a Dropout layer (0.2), the model proceeds through a Dense layer (64 units) with ReLU activation. The final output layer uses softmax activation, suitable for multi-class classification.
   - LSTM: starts with an embedding layer, followed by an LSTM layer (128 units) to capture sequential dependencies in the data. After a Dropout layer (0.2), the output is flattened and passed through a Dense layer (64 units) with ReLU ac-

tivation (Telmem et al., 2024). The final output layer uses softmax activation, classifying the data into one of the 5 output classes.

Following the completion of all experiments for binary classification, the Bi-RNN, Fusion Model, and LSTM+GRU Hybrid exhibited promising accuracy. These models were then submitted to the task organizer, where the Fusion Model attained the highest performance on the test set. In Task B, the three models outlined in the methodology were submitted, with the LSTM+GRU Hybrid emerging as the top performer among them. The results of the submitted models on the development set are presented in Tables 3 and 4. The proposed methodology is illustrated in Figure 1. The implementation code is available on GitHub.

4 Results

The results from the test set reveal differences in performance between Task A and Task B. In Task A, i.e., Fake vs Original news classification, the Fusion Model achieved a macro-average score of 0.792, demonstrating strong performance in distinguishing between fake and original news. However, in Task B (fine-grained fake news classification), the LSTM + GRU Hybrid model scored a much lower macro-average of 0.26, highlighting the difficulty of classifying nuanced categories like Half True, False, and Mostly False.

Model	Precision	Recall	F1-score
Bi-RNN	0.75	0.75	0.75
Fusion Model	0.83	0.82	0.82
LSTM-RNN Hybrid	0.81	0.79	0.79

Table 3: Performance of Task A on Development Set.

Table 3 summarizes the metrics, including accuracy and F1-scores, for Task A, while Table 4 outlines the development set performance for Task B. Test set results and Ranking for both tasks can be accessed on the Fake News Detection in Dravidian Languages DravidianLangTech@NAACL 2025 task page.

Model	Precision	Recall	F1-Score
GRU-Model	0.63	0.67	0.63
LSTM Model	0.58	0.63	0.60
LSTM + GRU Hybrid	0.94	0.93	0.93

Table 4: Performance of Task B on Development Set

In this study, multi-class classification (F1-score 0.26) is hindered by class imbalance and the inability to discern subtle categories. Because sequential models don't have a deep understanding of context, it's more difficult to spot sarcasm and implicit misinformation. In environments with limited resources, real-time deployment is limited by the high computational cost. Reliability is decreased when fact-checking procedures are absent. Interpretability and trust are impacted by deep learning models' black-box nature. Explainability strategies and attention-based models should be investigated in future research.

The low score in Task B may be due to class imbalance and semantic overlap between categories, making it harder for the model to distinguish subtle differences. While the hybrid model captures sequential patterns, it may lack the ability to encode deeper contextual cues, especially without attention mechanisms or transformer-based embeddings. Additionally, the macro-average F1-score penalizes poor performance in labels with minimal classes, further lowering the overall score. These results suggest the need for class balancing, more advanced embeddings like BERT or XLM-R, and attention mechanisms to improve performance.

5 Conclusion

This study examined hybrid models that focus on binary and multi-class classification for the detection of fake news in Dravidian languages. In binary classification, the Fusion Model obtained a robust macro-average F1-score of 0.792, whereas the LSTM + GRU hybrid model had trouble in multi-class classification, achieving an F1-score of 0.26. While highlighting the usefulness of hybrid models for binary tasks, these results also point to the need for more sophisticated techniques in multi-class classification. Attention mechanisms, transformers, context-aware embeddings, and language-specific preprocessing methods for Dravidian languages could all be included in future research.

References

Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. 2022. A comparative study of machine learning and deep learning techniques for fake news detection. *Information*, 13(12):576.

Saleh Alyahyan. 2025. Fusionnet remote a hybrid deep learning ensemble model for remote image classification in multispectral images. *Discover Computing*, 28(1):3.

Tavishee Chauhan and Hemant Palivela. 2021. Optimization and improvement of fake news detection using deep learning approaches for societal benefit. *International Journal of Information Management Data Insights*, 1(2):100051.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

MDP P Goonathilake and PPN V Kumara. 2020. Cnn, rnn-lstm based hybrid approach to detect state-of-the-art stance-based fake news on social media. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 23–28. IEEE.

Ramin Mousa, Mitra Khezli, Mohamadreza Azadi, Vahid Nikoofard, and Saba Hesaraki. 2024. Classifying objects in 3d point clouds using recurrent neural network: A gru lstm hybrid approach. *arXiv preprint arXiv:2403.05950*.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Meryam Telmem, Naouar Laaidi, Youssef Ghanou, Sanae Hamiane, and Hassan Satori. 2024. Comparative study of cnn, lstm and hybrid cnn-lstm model in amazigh speech recognition using spectrogram feature extraction and different gender and age dataset. *International Journal of Speech Technology*, 27(4):1121–1133.

Muhammad Umer, Zainab Imtiaz, Muhammad Ahmad, Michele Nappi, Carlo Medaglia, Gyu Sang Choi, and Arif Mehmood. 2023. Impact of convolutional neural network and fasttext embedding on text classification. *Multimedia Tools and Applications*, 82(4):5569–5585.

Muhammad Waqas and Usa Wannasingha Humphries. 2024. A critical review of rnn and lstm variants in hydrological time series predictions. *MethodsX*, page 102946.

Cong Xu. 2024. Cnn-gru model for ecg signal classification using ucr time series data. *Advances in Engineering Innovation*, 12:31–35.

Xinzhi Yang, Zili Fang, Wenbo Zhang, Lixia Xi, Xiaoguang Zhang, and Nan Cui. 2022. Fiber nonlinear compensation using bi-directional recurrent neural network model based on attention mechanism. In *2022 Asia Communications and Photonics Conference (ACP)*, pages 426–428. IEEE.

Mesay Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2024. Habesha@ dravidianlangtech 2024: Detecting fake news detection in dravidian languages using deep learning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 156–161.

Mesay Gemeda Yigezu, Girma Yohannis Bade, Olga Kolesnikova, Grigori Sidorov, and Alexander F Gelbukh. 2023. Multilingual hope speech detection using machine learning. In *IberLEF@ SEPLN*.

Girma@DravidianLangTech 2025: Detecting AI Generated Product Reviews

Girma Yohannis Bade[1,a], **Muhammad Tayyab Zamir**[2,a], **Olga Kolesnikova**, **José Luis Oropeza**[4,a], **Grigori Sidorov**[5,a], **Alexander Gelbukh**[6,a]
[a]Centro de Investigaciones en Computación(CIC),
Instituto Politécnico Nacional(IPN), Miguel Othon de Mendizabal,
Ciudad de México, 07320, México.
[1]girme2005@gmail.com

Abstract

The increasing prevalence of AI-generated content, including fake product reviews, poses significant challenges in maintaining authenticity and trust in e-commerce systems. While much work has focused on detecting such reviews in high-resource languages, limited attention has been given to low-resource languages like Malayalam and Tamil. This study aims to address this gap by developing a robust framework to identify AI-generated product reviews in these languages. We explore a BERT-based approach for this task. Our methodology involves fine-tuning a BERT-based model specifically on Malayalam and Tamil datasets. The experiments are conducted using labeled datasets that contain a mix of human-written and AI-generated reviews. Performance is evaluated using the macro F1 score. The results show that the BERT-based model achieved a macro F1 score of 0.6394 for Tamil and 0.8849 for Malayalam. Preliminary results indicate that the BERT-based model performs significantly better for Malayalam than for Tamil in terms of the average Macro F1 score, leveraging its ability to capture the complex linguistic features of these languages. Finally, we open the source code of the implementation in the GitHub repository: AI-Generated-Product-Review-Code.

Keywords: AI-generaated, Detection, Product review, BERT

1 Introduction

The e-commerce marketplaces has led to an increase in online product reviews, which have become important for consumer buying behavior. Nonetheless, the problem of AI content generation has worsened the trust issues surrounding platforms due to the flood of fake reviews. These types of review can deceive users, affect the company's image, and alter competitive structures, making it difficult to devise new methods to identify them (Ott et al., 2011; Banerjee and Chua, 2014). With so many efforts being made to identify fake reviews in the English language, not many focus is acquired towards low-resource languages such as Malayalam and Tamil. This gap needs to be filled as these low resource languages are unique in their own way.

NLP does not tend to focus much on Malayalam and Tamil, which are common in South India as well as with the diaspora because resources are limited (Joshi et al., 2020; Zamir et al., 2024a).

The goal for this particular research is to create an effective hybrid system for the detection and classification of AI-produced reviews in Malayalam and Tamil using modern transformer. In particular, we utilize a BERT-based model which has been fine tuned to Malayalam and Tamil datasets. To explain it in simple terms, BERT is a great performer in most NLP learning tasks because it uses context and other deep language features to understand text's meaning (Devlin, 2018; Ahani et al., 2024), and we tend to fine tune BERT to specific low resource languages datasets to help it perform even better.

The described work requires the collection, and careful labeling of approximately authentic-sounding reviews of AI technology in both languages, which have been written in a mix of human and AI text. Both training and evaluation of the models are performed using the Macro F1-score, and their changes. Initial analysis indicates that the BERT-based model ultimately achieves the best results for Malayalam language as compared to Tamil language baseline making it a great alternative for cases involving complex linguistic phenomena (Ullah et al., 2024).

2 Related Works

With the advanced evolution of AI-generated text models, such as GPT-4 and its future models, the content generation process has changed in several spheres, including reviews of online products. Al-

though there's an enormous opportunity in these tools, their inappropriate use has brought forth concerns regarding authenticity, especially in sensitive fields like consumer decision-making. The problem becomes more pronounced in low-resource languages like Malayalam and Tamil since there is still very little research done in an AI generated content detection. Other literature point towards resource rich languages, like English, and emphasize on the ethical consideration as well as the systems in place of text generation detection systems (Solaiman et al., 2019). Nevertheless, the application of these techniques in low resource languages poses a problem owing to the distinctive linguistic features of these languages.

Malayalam and Tamil belong to low resource languages that is characterized by its high morphology, agglutinative nature and sophisticated syntactic features. Research has demonstrated that such languages are a very difficult case for natural language processing (NLP) owing to suffixation, highly compound words and poor availability of standard transliteration (Annamalai, 2010; Chakravarthi et al., 2022b) which makes it hard to use many methods designed for resource-rich languages without major adaptations.

This is why such models cannot simply be transferred directly from resource-rich languages as they require specialized approaches to help, for example text classification or AI related content detection.

While datasets are the be-all and end-all for training detection models, Dravidian languages have short supply of annotated data. One example here is the "Dravidian CodeMix" shared task (Chakravarthi et al., 2023; Tash et al., 2024) which provides a dataset of code-mixed sentiment analysis and offensive language detection. Especially, there are few datasets being curated for AI-generated content detection in Dravidian languages. Thus, works such as "Shared Task for Detector AI-generated Product Reviews in Dravidian Languages" fill this need by providing training and test datasets in Malayalam as well Tamil, allowing researches to train their models suited for these languages.

Methods (when it comes to detecting AI-generated text) were primarily based on linguistic characteristics analysis and stylometry like n-gram comparation, detecting a style inconsistence (Jawahar et al., 2019). Recent transformer powered models like BERT, RoBERTa and mBERT have established excellent performance in text classification tasks due to the impact of deep learning. Generalizable Multilingual models (XLM-R (Jawahar et al., 2019; Chakravarthi et al., 2022a)) that outperforms in low-resource languages, are showing an encouraging performance when they are finetuned on domain-specific data. Performance of AI-based text detection models is evaluated based on evaluation metrics (especially when we only care about identifying the text generated by AI), and developing such models is no exception.

In NLP, the score of F1 is an accuracy metric that strikes a balance between recall and precision (Naidu et al., 2023), reflecting on the other hand it was suited for classification problems with imbalanced datasets (Bade et al., 2024c). Adoption of the metric in this Shared Task underscores the need for a more comprehensive model performance evaluation measure for low-resource languages (Priyadharshini et al., 2022; Zamir et al., 2024b). One of the biggest ethical concerns in identifying AI-generated content detection is within low resource languages, to keep trust on online platforms alive. Kimera et al. (2024) has claimed that systematic detection systems can help to prevent false information and building trustworthy digital eco-system. In future work, we will address issues of more diverse and representative datasets (including multimodal), as well as breaking biases to improve detection systems for these languages.

3 System Description

In this section, we discuss about datasets, preprocessing, feature extraction, and model seletion. Moreover, finally it overviews architecture of this task.

3.1 Datasets

The research in the NLP domain heavily relies on well-curated datasets, which serve as the driving force for creating intelligent systems (Bade et al., 2024a). However, it is labor intensive to obtain well-written data to train the language model, especially under-resourced ones (Bade, 2021). Thanks to DravidianLangTech (Priyadharshini et al., 2023), they offered datasets and task instructions for this work (Premjith et al., 2025). The datasets are organized into two subsets: training and test sets. The training data set is a data set that contains two variables, X input and Y output. While the X variable represents users' comments, the Y variable repre-

sents their values that can determine whether the comments are human written or machine generated. However, the test dataset does not contain the Y variable because we expect it to be predicted by the model. This kind of dataset arrangements are more conveinent for supervised machine learning. Mathematically,

$$\text{Training Data} = \sum_{j=1}^{n} (X_{ij}, Y_{ij})$$

and

$$\text{Test Data} = \sum_{j=1}^{m} X_{ij}$$

Table 1 presents the detailed statistics of these datasets.

Languages	Dataset	has_label?	Size
Tamil	Train	yes	808
	Test	no	100
	Total	**–**	**908**
Malayalam	Train	yes	800
	Test	no	200
	Total	**–**	**1,000**

Table 1: Dataset statistics

Table 1 outlines the training and test datasets. While training dataset served as the primary resource for training the selected algorithm, test dataset served to evaluate the final performance of the model. Notably, test data was kept separate and remains unseen during the training process. Table 2 further provides the class label distribution for the training.

Language	Label	Count
Tamil	HUMAN	403
	AI	405
	Total	**808**
Malayalam	HUMAN	400
	AI	400
	Total	**800**

Table 2: The statistics of class label distributions of the training dataset.

3.2 Preprocessing

The annotated training, development datasets, and test dataset underwent pre-processing. Then punctuation mark removal, emoji removal, and username removal are the main objectives of this step in this particular use case. The built-in "re" module in Python has helped to eliminate all these staff.

3.3 Feature Extraction

Since AI algorithms operate on numeric data (Bade and Seid, 2018), it is necessary to encode the input of text to a numeric equivalent (Bade et al., 2024d). The process of converting text input into numeric form is known as data encoding or feature extraction (Bade et al., 2024a). This task is carried out by BertTokenizer of the BERT model.

3.4 Model Selection

Once the NLP processing steps such as dataset organization, pre-processing, and feature extraction are completed, the next critical step involves selecting and applying appropriate AI algorithms. In this study, we employ the BERT model, a state-of-the-art architecture based on Transformers, to achieve our objectives (Yigezu et al., 2023), specifically the bert-base-uncased variant. Since its introduction, Transformers have revolutionized NLP by enabling enhanced parallelization and effectively capturing long-range dependencies (Vaswani, 2017). Among these models, BERT remains a fundamental baseline, consistently achieving state-of-the-art performance across various NLP benchmarks (Rogers et al., 2021).

To process textual data, BERT employs its dedicated BertTokenizer, which tokenizes text and converts it into numerical representations, ensuring efficient input processing for the model (Bade et al., 2024b). Table 3 outlines the hyperparameters used for this model.

Hyperparameters	Values
Learning Rate	1e-5
Evaluation Strategy	Epoch
Epochs	5
Batch Size	32
Activation function@output level	Sigmoid

Table 3: BERT Hyperparameters

As we can see from Table 3, the learning rate indicates the number of times the exectuion taken place to improve the model's performance. Epoch refers to one complete pass through the entire training dataset by the learning algorithm (Mersha et al., 2024). Thus, we set the epoch to be 5,.i.e the execution did pass 5 complete times. The batch size refers to dividing the total data size into 32 and

bringing the divided batch one a time for the execution. This helps the execution to be fast. The last parameter, activation function is used to label or group the computational result into two classes. Figure 1 presents the workflow of this study.

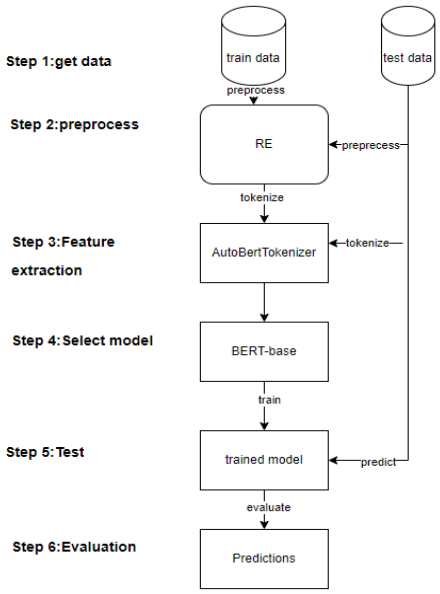


Figure 1: Overall workflow of the study.

In Figure 1, the pipeline begins with acquiring a training dataset, followed by pre-processing and feature extraction steps. The selected model, BERT-base, is then trained to learn the patterns and behavior specific to this dataset. Once training is complete, the model is evaluated using test data to generate predictions.

4 Result and Discussion

We trained the provided dataset on the BERT model and tested its performance using a separate test set. The predictions generated from the test set data were submitted to the workshop organizers for evaluation. These submissions were evaluated using accuracy (Acc), average macro precision (P), average macro recall (R), and average macro F1-score (F1) as evaluation metric. The final results, published by the organizers, revealed that for Tamil, BERT attained a significantly higher macro F1 score of 0.6394, and for Malayalam, it excelled with a macro F1 score of 0.8849. Table 4 shows more details.

From Table 4, we can easily infer that the selected model and configured hyperparameters are more favored for Malayalam than Tamil.

Language	Acc	P	R	F1
Tamil	0.6400	0.6394	0.6394	0.6394
Malayalam	0.8850	0.8859	0.8850	0.8849

Table 4: Performance metrics of **BERT** model for the data of Malayalam and Tamil languages.

5 Comparative Analysis

The suggested AI-generated review detection model is contrasted with other baseline techniques, such as naive Bayes, support vector machines (SVMs), BERT, ALBERT, RoBERTa, and gradient boosting decision trees (GBDTs).These baseline methods are compared to our approaches in terms of the F1 measure, regardless of the dataset they employed.

Model	F1-score
Qualitative (Fröhnel et al., 2025)	0.5300
GANs(Ke et al., 2025)	0.9500
RoBERTa (Wang et al., 2025)	0.7342
BERT (**Ours**)Mal	**0.8849**
BERT (**Ours**)Tam	0.6394

Table 5: Comparison of the models of our work and others. The result in bold shows the performance achieved by our approach, revealing the effectiveness of the model.

6 Conclusion and Future Work

This research created a methodology for detecting reviews generated by AI for products in Malayalam and Tamil using a fine-tuned BERT model. The experimental results proved that the BERT model is the best performer and scored 0.8849 on Macro F1-score for the Malayalam language compared to 0.6394 scored by the Tamil language. The findings support the claim regarding the sophisticated grammatical features possessed by BERT for these low resource languages, which makes the translation of these languages into other languages rather appealing due to the challenges posed by the insufficient resources. This work highlights the fact that the models for advanced languages are necessary for the impoverished context and attempts to provide tools that could be used to improve the language's authenticity.

In the future, this work will be improved both by adding more diverse domains and by using multilingual models in potential cross-lingual transfer learning setups. BERT can also be enhanced by

using explainable AI techniques. This will help to increase the accuracy of detection and support broader use cases in combating AI-generated content.

Limitation and Ethics Statement

The datasets for Tamil and Malayalam languages used in this study were limited in size, and the model was trained on this relatively small dataset. As a result, the observed performance may not generalize well to all unseen data. Despite these constraints, our model demonstrated comparable performance in detecting AI-generated product reviews for social media posts. Nevertheless, in a highly competitive environment, our method achieved impressive rankings of 10th and 32nd for Malayalam and Tamil, respectively. Additionally, our work adheres to the ethical principles outlined for computational research and professional conduct[1].

Acknowledgment

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Z Ahani, M Tash, M Zamir, and I Gelbukh. 2024. Zavira@ dravidianlangtech 2024: Telugu hate speech detection using lstm. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 107–112.

Elay Annamalai. 2010. Politics of language in india. In *Routledge Handbook of South Asian Politics*, pages 213–231. Routledge.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024a. Social media hate and offensive speech detection using machine learning method. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 240–244.

Girma Yohannis Bade. 2021. Natural language processing and its challenges on omotic language group of ethiopia. *Journal of Computer Science Research*, 3(4):26–30.

Girma Yohannis Bade, O Koleniskova, José Luis Oropeza, Grigori Sidorov, and Kidist Feleke Bergene. 2024b. Hope speech in social media texts using transformer. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS. org*.

Girma Yohannis Bade, Olga Kolesnikova, and Jose Luis Oropeza. 2024c. Evaluating the quality of data: Case of sarcasm dataset.

Girma Yohannis Bade, Olga Kolesnikova, José Luis Oropeza, and Grigori Sidorov. 2024d. Lexicon-based language relatedness analysis. *Procedia Computer Science*, 244:268–277.

Girma Yohannis Bade and Hussien Seid. 2018. Development of longest-match based stemmer for texts of wolaita language. *vol*, 4:79–83.

Snehasish Banerjee and Alton YK Chua. 2014. Applauses in hotel reviews: Genuine or deceptive? In *2014 Science and Information Conference*, pages 938–942. IEEE.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Cn, Sangeetha S, Malliga Subramanian, Kogilavani Shanmugavadivel, Parameswari Krishnamurthy, Adeep Hande, Siddhanth U Hegde, Roshan Nayak, and Swetha Valli. 2022a. Findings of the shared task on multi-task learning in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 286–291, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[1] https://www.aclweb.org/portal/content/acl-code-ethics

Kim Fröhnel, Bennet Santelmann, and Rüdiger Zarnekow. 2025. Genuine or fake? explaining consumers' perception and detection of ai-generated fake reviews.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Zong Ke, Shicheng Zhou, Yining Zhou, Chia Hong Chang, and Rong Zhang. 2025. Detection of ai deepfake and fraud in online payments using gan-based models. *arXiv preprint arXiv:2501.07033*.

Richard Kimera, Yun-Seon Kim, and Heeyoul Choi. 2024. Advancing ai with integrity: Ethical challenges and solutions in neural machine translation. *arXiv preprint arXiv:2404.01070*.

Melkamu Abay Mersha, Girma Yohannis Bade, Jugal Kalita, Olga Kolesnikova, Alexander Gelbukh, et al. 2024. Ethio-fake: Cutting-edge approaches to combat fake news in under-resourced languages using explainable ai. *Procedia Computer Science*, 244:133–142.

Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. 2023. A review of evaluation metrics in machine learning algorithms. In *Computer Science On-line Conference*, pages 15–25. Springer.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

M Tash, Z Ahani, M Zamir, O Kolesnikova, and G Sidorov. 2024. Lidoma@ lt-edi 2024: Tamil hate speech detection in migration discourse. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 184–189.

Fida Ullah, Muhammad Zamir, Muhammad Arif, M Ahmad, E Felipe-Riveron, and Alexander Gelbukh. 2024. Fida@ dravidianlangtech 2024: A novel approach to hate speech detection using distilbert-base-multilingual-cased. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 85–90.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, et al. 2025. Genai content detection task 1: English and multilingual machine-generated text detection: Ai vs. human. *arXiv preprint arXiv:2501.11012*.

Mesay Gemeda Yigezu, Girma Yohannis Bade, Atnafu Lambebo Tonja, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Bilingual word-level language identification for omotic languages. In *International Conference on Advances of Science and Technology*, pages 63–77. Springer.

M Zamir, M Tash, Z Ahani, A Gelbukh, and G Sidorov. 2024a. Tayyab@ dravidianlangtech 2024: detecting fake news in malayalam lstm approach and challenges. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 113–118.

Muhammad Tayyab Zamir, Fida Ullah, Rasikh Tariq, Waqas Haider Bangyal, Muhammad Arif, and Alexander Gelbukh. 2024b. Machine and deep learning algorithms for sentiment analysis during covid-19: A vision to create fake news resistant society. *PloS one*, 19(12):e0315407.

Beyond_Tech@DravidianLangTech 2025: Political Multiclass Sentiment Analysis using Machine Learning and Neural Network

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2], **Sanjai R**[1], **Mohammed Sameer B**[1], **Motheeswaran K**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{sanjair.22aid, mohammedsameerb.22aid}@kongu.edu
motheeswarank.22aid@kongu.edu

Abstract

Research on political feeling is essential for comprehending public opinion in the digital age, as social media and news platforms are often the sites of discussions. To categorize political remarks into sentiments like positive, negative, neutral, opinionated, substantiated, and sarcastic, this study offers a multiclass sentiment analysis approach. We trained models, such as Random Forest and a Feedforward Neural Network, after preprocessing and feature extraction from a large dataset of political texts using Natural Language Processing approaches. The Random Forest model, which was great at identifying more complex attitudes like sarcasm and opinionated utterances, had the greatest accuracy of 84%, followed closely by the Feedforward Neural Network model, which had 83%. These results highlight how well political discourse can be analyzed by combining deep learning and traditional machine learning techniques. There is also room for improvement by adding external metadata and using sophisticated models like BERT for better sentiment classification.

1 Introduction

Political sentiment analysis is crucial for understanding public opinions on political topics, particularly in today's digital age where political discussions have largely moved to online platforms like social media. These discussions generate vast amounts of textual data, but political sentiment is complex, ranging from neutral and supportive to oppositional and sarcastic, making analysis challenging.

In this study Chakravarthi et al. (2025), we propose a multiclass sentiment analysis method for political text, categorizing sentiments into positive, negative, neutral, substantiated, opinionated, and sarcastic. Using NLP techniques such as tokenization, lemmatization, and TF-IDF for feature extraction, we compare the performance of machine learning models, including Random Forest and Neural Networks. Neural Networks performed better, particularly in identifying subtle emotions like sarcasm. Our results show that combining traditional machine learning with deep learning improves sentiment analysis accuracy in political discourse, despite ongoing challenges with sarcasm detection.

2 Related Works

The paper Ma'Aly et al. (2024) examined the multi-label sentiment categorization of YouTube comments from the 2024 Indonesian presidential election using CNN, Bi-LSTM, and hybrid CNN-BiLSTM models. The model that captured long-term dependencies, the Bi-LSTM, had the best accuracy of 98% and the highest AUC of 0.92. In order to address class imbalance, preprocessing techniques included normalization, stopword removal, text augmentation, and class weights. Tun and Khaing (2023) development of superior sentiment lexicons for political tweets is the main subject of this work, which investigates Twitter's function in political discourse and sentiment analysis. In sentiment classification, the Linear SVC model achieved 98% accuracy, outperforming the Multinomial Naive Bayes (MNB) and Decision Tree (DT) models.To enhance tweet quality, the study (Saranya and Usha, 2023) suggests a sentiment analysis technique utilizing TF-IDF and Intelligent WordNet lemmatization. Emotion detection using a Random Forest network outperforms current multiclass sentiment classification methods with an accuracy of 90%. The method goes beyond merely detecting emotions and places a strong emphasis on closely examining tweets that are positive or negative.

Mu et al. (2024) presented a model for multimodal sentiment analysis of government comments

using a unique cross-attention fusion network and contrastive learning, which achieves 96.80% accuracy. Policymakers benefit from improved emotion polarity recognition thanks to the model's 10.21% accuracy gain over the CLIP model. Diqi et al. (2024) used the Multinomial Naive Bayes algorithm to analyze sentiment in digital election campaign ads with 96% accuracy. It highlights the necessity of customized approaches and suggests that future research investigate cutting-edge NLP methods using real-time social media data. Innork et al. (2023) investigated a number of machine learning techniques for sentiment analysis, such as Random Forest, K-Nearest Neighbors, Support Vector Machine, Multinomial Naive Bayes, and an ensemble method. According to the study, when it comes to categorizing hotel evaluations, the ensemble approach works better than the others.

Kowsik et al. (2024) used sentiment analysis to classify political leanings in Twitter data, the study "Sentiment Analysis of Twitter Data to Detect and Predict Political Leniency Using Natural Language Processing" achieves a 99% confidence level in identifying the political biases of user profiles .In order to forecast the results of legislative debates, Salah (2014) investigates sentiment analysis techniques based on classification and vocabulary. It presents the Debate Graph Extraction framework and suggests domain-specific sentiment lexicons to display and examine the sentiment and structure of disputes. Liebeskind et al. (2017) examined 5.3 million Facebook comments about politicians and evaluates nine machine learning techniques for sentiment analysis. Classifying generic attitudes against content-specific attitudes revealed differences, with n-gram representation being the most successful and logistic regression achieving the best accuracy.

Sumathy and Muthukumari (2018) focused on sentiment analysis of social media reviews using machine learning. A Support Vector Machine (SVM) was used to classify reviews as positive or negative, outperforming Naive Bayes in terms of accuracy. The SVM model was optimized with a multi-class kernel and hyperparameter tuning.

3 Problem Description

Sentiment analysis of Tamil political debates is difficult because of the language's complexity, cultural quirks, and emotions like positive, negative, neutral, opinionated, substantiated, and ironic. Classifier development is made more challenging by the absence of annotated datasets, and traditional binary sentiment analysis is insufficient. This work creates a multiclass sentiment analysis model for Tamil political literature in order to handle problems such as inconsistent data, language normalization, and the intricacy of political speech. Prediction, model training, feature extraction, and data preparation for multiclass sentiment classification are all included in the system. The dataset used is provided by the codalab Chakravarthi et al. (2025).Our system performed competitively, placing 16th Rank among 153 participants in this shared task.

3.1 Data Preprocessing

In sentiment analysis, preprocessing is essential, especially when using a regional language like Tamil. Several processes are used to clean and normalize the raw textual data:

- Normalization of Unicode Characters: To guarantee consistent text representation, unwanted Unicode characters are eliminated.
- Elimination of Non-Tamil Characters: Only the pertinent script is kept after all non-Tamil characters, special symbols, and numerical values are filtered out.
- Handling Spoken Variants: The approach normalizes vowels and diphthongs to their regular written forms to accommodate typical spoken variants in Tamil.
- Tokenization: Stanza is used to tokenize the preprocessed text, which aids in breaking it into discrete words or tokens.
- The consistency and cleanliness of the textual data are guaranteed by this pre-processing pipeline, preparing it for feature extraction.

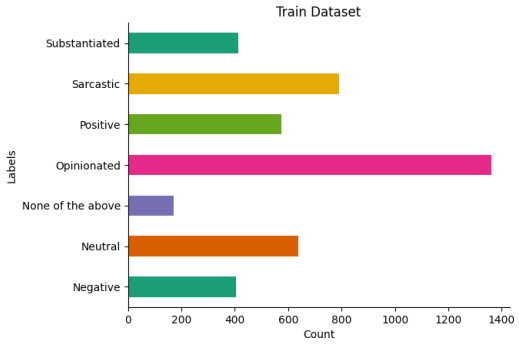


Figure 1: Train dataset Labels and it's count

3.2 Feature Extraction

Following preprocessing, we use the TF-IDF (Term Frequency-Inverse Document Frequency) approach

to extract features from the text. Each word's significance in relation to the overall dataset is captured by TF-IDF, which transforms the text into numerical vectors. The model is able to concentrate on important phrases that are essential for differentiating various feelings thanks to this modification.

3.3 Balancing the Dataset

Sentiment classification datasets frequently have unequal class representation, with certain sentiment categories (like neutral) being overrepresented and others (like sarcasm) being underrepresented. The system uses to upsampling for the minority classes in order to lessen this problem. In order to guarantee that the dataset is balanced and that every sentiment class is equally represented throughout training, this technique creates extra samples for the underrepresented classes.

4 Methodology

In this study, two machine learning algorithms are used to categorize Tamil political literature into six different sentiment categories: sarcastic, opinionated, substantiated, neutral, positive, and negative.

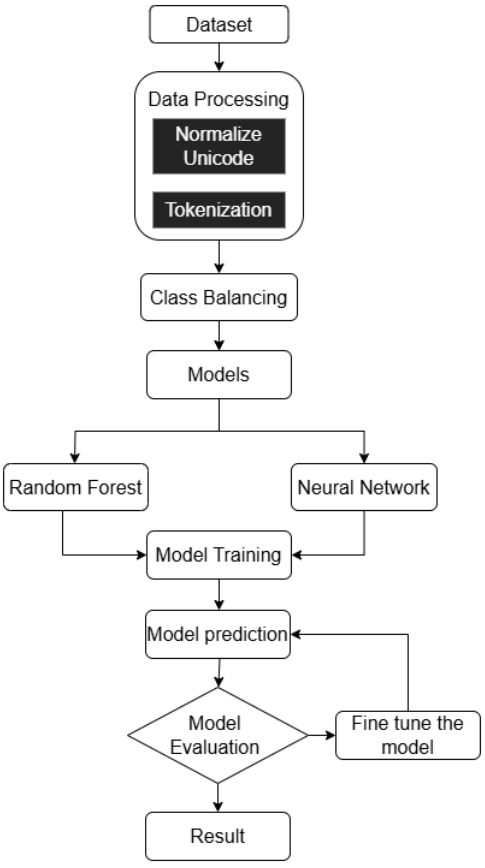


Figure 2: Proposed System Workflow

The models were trained to predict these sentiment classes after the data was preprocessed using Natural Language Processing (NLP) techniques and features were extracted using TF-IDF.

4.1 Random Forest Classifier

An ensemble learning method called the Random Forest algorithm creates several decision trees during training and averages their predictions to increase classification accuracy. This approach is reliable because it uses the combined output of multiple independent trees to minimize overfitting.

- Model Input: The TF-IDF feature vectors produced from the preprocessed Tamil text data were used to train the Random Forest model. The significance of each word in the text is represented by these vectors.
- Model Configuration: We set up 100 decision trees (n_estimators=100) in the Random Forest classifier. A random selection of characteristics was used to train each tree, and hyperparameters like each tree's depth were adjusted to balance the effectiveness and performance of the model.
- **Performance:** Random Forest achieved 95% accuracy on training data and 84% on test data. It performed well in classifying positive, neutral, and negative sentiments. However, it struggled with nuanced emotions like sarcasm and well-supported claims.

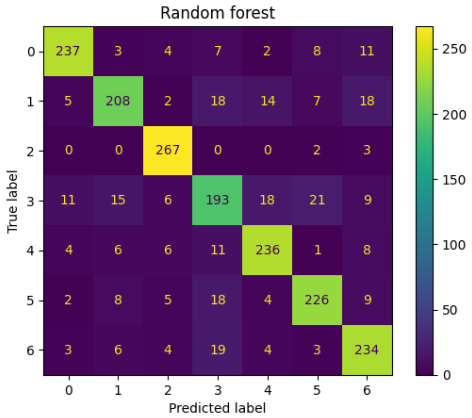


Figure 3: Confusion Matrix for Random Forest

Metric	Precision	Recall	F1-Score
Accuracy	0.84	-	-
Macro avg	0.84	0.84	0.84
Weighted avg	0.84	0.84	0.84

Table 1: Classification Report for Random Forest Model

4.2 Feedforward Neural Network

The Random Forest model might not be able to properly handle the deeper and more complex patterns in the text, so a Feedforward Neural Network model was used. Because of their superior ability to learn non-linear correlations, neural networks are well-suited for the complex task of classifying political mood.

1. Architecture for the Model: Several layers made up the neural network architecture:
    - Input Layer: The TF-IDF vectorized text data was sent to the input layer.
    - Four thick hidden layers of 512, 256, 128 and 64 neurons each were included in the model. To add non-linearity and enable the model to learn intricate correlations between features, each hidden layer employed the ReLU activation function.
    - Dropout Layers: By randomly deactivating neurons during training, dropout layers were added to minimize overfitting.
    - Output Layer: Using a softmax activation function, the output layer categorized the text into six sentiment categories: sarcastic, opinionated, substantiated, neutral, positive, and negative.

2. Model Training: The Adam optimizer with a sparse categorical cross-entropy loss function was used to train the neural network. With a batch size of 128 and 30 epochs, the model was trained with the goal of maximizing classification accuracy while minimizing loss.

3. Performance: Training with neural network the accuracy was 94% and test data accuracy was 83%. It was marginally less successful than the Random Forest model in categorizing neutral sentiments, but it did well in detecting subtle sentiments like sarcasm and opinionated utterances.

Metric	Precision	Recall	F1-Score
Accuracy	0.84	-	-
Macro avg	0.83	0.84	0.83
Weighted avg	0.83	0.84	0.83

Table 2: Classification Report for Neural Network Model

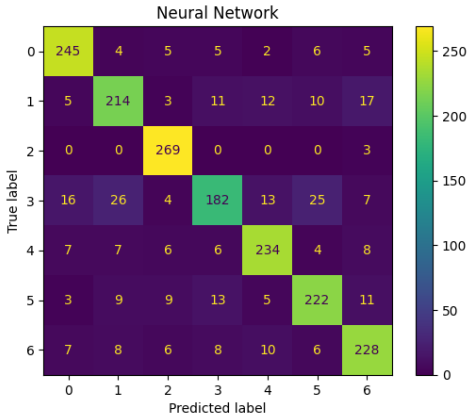


Figure 4: Confusion Matrix for Neural Network model

5 Conclusion

In this experiment Sanjai (2025), we engineered and tested two machine learning classifiers: Random Forest and a Neural Network, toward multiclass sentiment analysis of the Tamil political corpus, with six classes: sarcasm, opinion, substantiation, neutral, positive, and negative. Aims were: to analyze these complex emotional varieties in political talk, which normally involve subtle nuances such as irony and opinions sustained by evidence. It is found that the Random Forest model has an accuracy of 84% and did well with the simpler sentiments, such as positive and neutral but failed to recognize more subtle sentiments like sarcasm and substantiated claims. The Neural Network performed better even though it exhibited a slightly lower overall accuracy of about 83% but picked up more of these intricate patterns and complex sentiments better on account of its deep learning architecture capturing latent features in text. That being said, the complementarity between these models suggests that a hybrid approach that utilizes traditional machine learning techniques along with deep learning might offer a more robust solution for political sentiment analysis. In the future, further studies may concern more sophisticated models, such as BERT, using context-aware representations with even greater improvement in sentiment classification accuracy, especially in capturing subtle emotional cues in political discussions. This can contribute toward better analysis of political texts and understanding of the nuances involved in public sentiments, ultimately contributing toward better political strategies and decision making.

References

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Mohammad Diqi, Dian Rhesa Rahmayanti, Marselina Endah Hiswati, I Wayan Ordiyasa, and Ida Hafizah. 2024. Digital democracy: Analyzing political sentiments through multinomial naive bayes in election campaign ads. *Jurnal Sistem Cerdas*, 7(2):237–247.

Kiatnumchai Innork, Jantima Polpinij, Umaporn Saisangchan, Manasawee Kaenumpornpan, Rungtip Charoensak, Theeraya Uttha, Khanista Namee, Bancha Luaphol, and Ajeej Meny. 2023. Machine learning-based multiclass classification methods for sentiment analysis. In *2023 7th International Conference on Information Technology (InCIT)*, pages 70–74. IEEE.

VV Sai Kowsik, L Yashwanth, Srivatsan Harish, A Kishore, Arun Cyril Jose, and Dhanyamol M V. 2024. Sentiment analysis of twitter data to detect and predict political leniency using natural language processing. *Journal of Intelligent Information Systems*, pages 1–21.

Chaya Liebeskind, Karine Nahon, Yaakov HaCohen-Kerner, and Yotam Manor. 2017. Comparing sentiment analysis models to classify attitudes of political comments on facebook (november 2016). *Polibits*, 55:17–23.

Ahmad Nahid Ma'Aly, Dita Pramesti, and Hanif Fakhrurroja. 2024. Comparative analysis of deep learning models for multi-label sentiment classification of 2024 presidential election comments. In *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, pages 502–507. IEEE.

Guangyu Mu, Chuanzhi Chen, Xiurong Li, Jiaxue Li, Xiaoqing Ju, and Jiaxiu Dai. 2024. Multimodal sentiment analysis of government information comments based on contrastive learning and cross-attention fusion networks. *IEEE Access*.

Zaher Salah. 2014. *Machine learning and sentiment analysis approaches for the analysis of Parliamentary debates*. Ph.D. thesis, University of Liverpool.

Sanjai. 2025. Political multiclass sentiment analysis.

S Saranya and G Usha. 2023. A machine learning-based technique with intelligentwordnet lemmatize for twitter sentiment analysis. *Intelligent Automation & Soft Computing*, 36(1).

P Sumathy and SM Muthukumari. 2018. Sentiment analysis of twitter data using multi class semantic approach. *2018 International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 3(6).

Yin Min Tun and Myo Khaing. 2023. A large-scale sentiment analysis using political tweets. *International Journal of Electrical & Computer Engineering (2088-8708)*, 13(6).

HTMS@DravidianLangTech 2025: Fusing TF-IDF and BERT with Dimensionality Reduction for Abusive Language Detection in Tamil and Malayalam

Bachu Naga Sri Harini[1], Kankipati Venkata Meghana [1], Kondakindi Supriya[1], S Tara Samiksha [1] Premjith B[1],
[1]Amrita School of Artificial Intelligence, Coimbatore,
Amrita Vishwa Vidyapeetham, India,
{cb.sc.u4aie24010,cb.sc.u4aie24022,cb.sc.u4aie24025}@cb.students.amrita.edu,
cb.sc.u4aie24045@cb.students.amrita.edu, b_premjith@cb.amrita.edu

Abstract

Detecting abusive and similarly toxic content posted on a social media platform is challenging due to the complexities of the language, data imbalance, and the code-mixed nature of the text. In this paper, we present our submissions for the shared task on abusive Tamil and Malayalam texts targeting women on social media—DravidianLangTech@NAACL 2025. We propose a hybrid embedding model that integrates embeddings generated using term frequency-inverse document frequency (TF-IDF) and BERT. To get rid of the differences in the embedding dimensions, we used a dimensionality reduction method with TF-IDF embedding. We submitted two more runs to the shared task, which involve a model based on TF-IDF embedding and another based on BERT-based embedding. The code for the submissions is available at https://github.com/Tarrruh/NLP_HTMS.

1 Introduction

Social media platforms are prevalent, and each of them has its own properties. People use these platforms to articulate their opinions on various topics. YouTube is a popular platform, which has a limited set of rules and regulations to control the comments posted by the users for videos. More freedom in expressing the ideas allowed the users to comment and interact using toxic content (Chakravarthi et al., 2023; Kavitha et al., 2020). Abusive language is one of the ways to spread toxic content, and it affects the mental well-being of others. Therefore, it is critical to maintain a healthy online environment.

Detecting abusive content from comments posted on social media and related interactions is a challenging task, mostly due to the properties of the language (Justen et al., 2022). Imbalance in data (Muzakir et al., 2022) and the presence of multilingual and code-mixed data make the detection more challenging (Kogilavani et al., 2023; Aporna et al., 2022). Researchers have utilised diverse methodologies to identify and alleviate harmful information on social media platforms. Transformer-based models, especially those based on BERT, have become extensively utilised for this task (Kalraa et al., 2021). Additionally, other deep learning algorithms like recurrent neural networks (RNNs) and their variants have gained widespread use (Mahmud et al., 2024; Darmawan et al., 2023).

In this paper, we discuss our submission to the shared task on abusive Tamil and Malayalam text targeting women on social media—DravidianLangTech@NAACL 2025 (Rajiakodi et al., 2025). We proposed a machine learning model by fusing embeddings generated using term frequency-inverse document frequency (TF-IDF) and BERT. To balance the dimensionality of TF-IDF and BERT embeddings, we reduced the TF-IDF embeddings, which are generally very high in dimension, to lower dimension using the random kitchen sink (RKS) algorithm (Sathyan et al., 2018).

The paper is structured outlined as follows. Section 2 reviews pertinent literature, Section 3 delineates the datasets employed, Section 4 elucidates the methodology, Section 5 discusses the analysis of experiments and results, and Section 6 concludes by summarizing findings.

2 Literature Review

Various machine learning and deep learning methods were developed to determine whether a comment in Dravidian language texts contains abusive content or not. In (Prasanth et al., 2022), the authors presented a Support Vector Machine (SVM) model that used a feature vector made with Term Frequency-Inverse Document Frequency (TF-IDF) along with character-level analysis and the Random Kitchen Sink (RKS) algorithm. The authors of (Subramanian et al., 2023) proposed adapter-

based multilingual transformer models based on MuRIL, XLM-RoBERTa, and mBERT to classify abusive comments in Tamil. The authors developed both conventional fine-tuned and adapter-based versions of the aforementioned models for classification. They observed that MuRIL had outperformed other models in this task with a detection accuracy of 74.7%. In (Chakravarthi et al., 2023), the authors employed machine learning algorithms such as naive Bayes, SVM, decision trees, random forests, and logistic regression with TF-IDF, Bag of Words (BoW), and FastText features to detect abusive comments in low-resource languages. In addition, the authors proposed deep learning models such as BiLSTM, BiLSTM with attention, mBERT, and XLM for this task. (Priyadharshini et al., 2022a) reported the submissions of the participants of the shared task. The participants used various machine learning and deep learning models for this task. The machine learning models include logistic regression, SVM, gradient boost classifiers, K-nearest neighbours, and ensemble models. Multi-layered perceptron (MLP), recurrent neural network (RNN), and Long Short-Term Memory (LSTM) are the examples of deep learning models submitted by the participants. In addition, the efficacy of transformer models such as mBERT, MuRIL, and XLM-R0BERTa was also investigated. The paper (Priyadharshini et al., 2022a) describes the systems developed for detecting abusive comments in Tamil, Tamil-English, and Telugu-English data. The authors reported that the majority of the systems used BERT-based models for feature extraction. There are models based on LSTM and traditional machine learning classifiers using TF-IDF features and word2vec embeddings. (Priyadharshini et al., 2023) and (Priyadharshini et al., 2022b) discuss different machine learning and deep learning models developed for detecting abusive content in Dravidian languages.

3 Dataset Description

The training dataset consists of the classification of abusive and non-abusive comments in Tamil and Malayalam. The organisers provided not only the training data, but also development (validation) data and test data without labels. We combined the training and development sets to train the model. Table 1 describes the dataset used for building the model.

Data Split	Number of Data Points	
	Tamil	**Malayalam**
Train	2,790	2,933
Validation	598	629
Test	598	629

Table 1: Dataset statistics showing the number of data points in each split

4 Methodology

In this work, we experimented with three models. These three models differ in the embeddings used. Following are the models we considered.

1. **TF-IDF embedding-based method**: In this model, we used TF-IDF embeddings. Here, we considered a maximum of 5000 unique words in the feature set. Therefore, the embedding model considered 3,000 most frequently occurring words for the study. We used the extracted features to train the classifier model, which we built using logistic regression. This is our Run 3 submission.

2. **A fused TF-IDF and BERT embeddings**: In this model, we computed both TF-IDF and BERT embeddings for the input text. We restricted the number of feature words to 5,000 based on their importance, similar to Run 3. TF-IDF learns the importance of a word in a sentence, whereas BERT generates an embedding by considering the context of the words. We generated the BERT embeddings using the 'bert-base-uncased' model. Because the TF-IDF and BERT embeddings have very different dimensions, the random kitchen sink (RKS) algorithm was used to reduce the size of the TF-IDF embeddings so that they are the same size as the BERT embeddings. RKS is a feature-mapping technique based on the Radial Basis Function that turns data into a space with desired dimensions. We concatenated these embeddings to obtain the representation of the input text. Machine learning classifiers were employed for training the models using these embeddings. This is our Run 2 submission.

3. **BERT-based method**: Here, we used the 'bert-base-uncased' model to generate the embeddings of the next text, which was further

fed into the machine learning classifiers to train the classification model. This is our Run 1 submission.

The flow-diagram of the model is illustrated in Figure 1.

5 Experimentation and Results

Based on the methodology, we divided the experiments into three.

5.1 TF-IDF embedding-based method

In this method, we used TF-IDF embeddings as feature representations. The TF-IDF algorithm converts text data into embeddings based on word frequency while reducing the influence of commonly occurring words. To control complexity and computational cost, we limit the number of features to 5,000 (max features = 5,000). We performed a 5-fold cross-validation (n splits = 5) with shuffling enabled (shuffle = True) to ensure a robust and reproducible evaluation of the model. The TF-IDF embeddings of the training data were used to train a logistic regression classifier. The maximum number of iterations was set to 1000 to ensure convergence, and the random state was kept constant to get stable results.

5.2 A fused TF-IDF and BERT embeddings

In this method, for feature extraction, we employed two methods: TF-IDF and BERT embeddings. We used the RKS algorithm with a gamma value of 1.0 to reduce the dimensionality of the TF-IDF embeddings from 5,000 features to 512. Simultaneously, we used the pre-trained bert-base-uncased model to generate contextual sentence embeddings. To get contextual representations, the input text was broken up into tokens that could be up to 512 characters long and processed in batches of 32. This was done on the T4 GPU of Google Colab. We obtained the final sentence embeddings by averaging the token representations from the last hidden layer. We then fused these two embeddings by concatenating the reduced TF-IDF and BERT embeddings horizontally. To make sure of the model's generalization capability, we did a 5-fold cross-validation with data. We trained a random forest classifier on the fused embeddings, using default hyperparameters.

5.3 BERT-based method:

We used deep contextual representations from the pre-trained BERT model, bert-base-uncased (with 768 dimensions), in this method. The input comments were processed to generate sentence embeddings using the BERT tokenizer, with truncation and padding applied for a maximum sequence length of 512 tokens. We processed the data efficiently with a batch size of 32 and passed the text through the BERT model. We computed the mean of the token representations from the last hidden layer to obtain a fixed-size embedding for each sentence. To train the Random Forest classifier, we used a 5-fold cross-validation method that included stratified K-Fold with shuffling and a random seed of 42 to make sure the results would be the same every time. The classifier was initialized with 100 decision trees (n estimators) in the forest and a maximum tree depth set to None, ensuring that the trees grew until all leaves were pure (or contained less than the minimum number of samples, which was set to 1).

5.4 Results

The training of the model with BERT embeddings and a random forest classifier demonstrated the potential for deep contextualisation from the BERT model. The performance of this approach varied depending on the dataset and the fold in the cross-validation. The BERT embeddings helped the model understand the hidden meanings of words and their context, which made it better at dealing with complicated language structures and subtle textual relationships. However, the computational cost of BERT made this method slower, particularly for larger datasets.

When combined with logistic regression, the TF-IDF approach trained significantly faster and required less computational power. Statistically, this method obtained the best results for the training set (almost the same as the method that used fused embedding).

The approach, which combined TF-IDF and BERT embeddings, yielded the most promising results by leveraging the strengths of both methods. TF-IDF captured word frequency and distribution information, while BERT provided semantic and contextual understanding. The fused embeddings allowed the model to capture both shallow and deep linguistic features, leading to more accurate predictions. This method performed well compared to

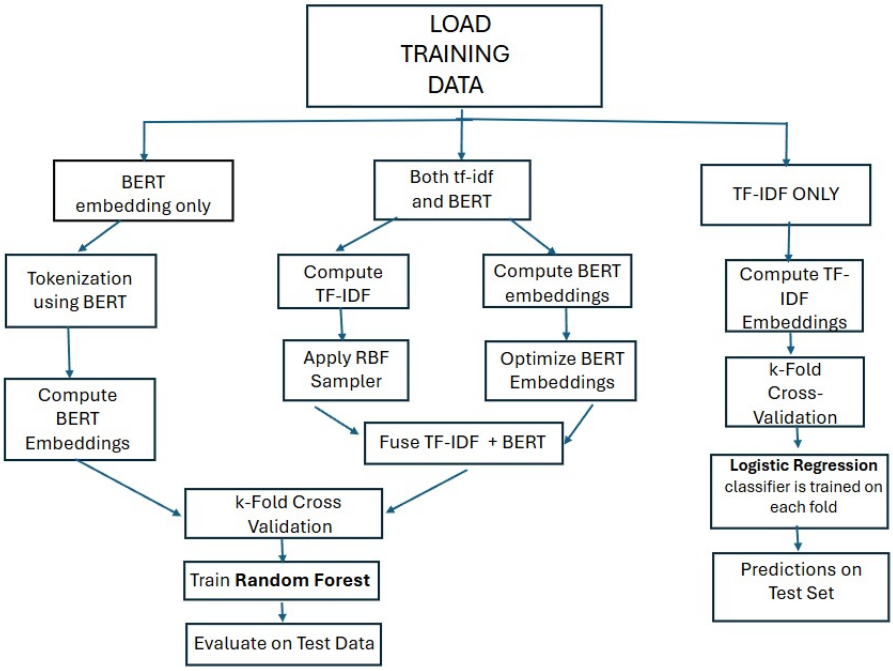


Figure 1: Block diagram explaining proposed methodology used for the task

Model	Precision	Recall	F1-score
TF-IDF	0.62	0.63	0.62
BERT	0.59	0.59	0.59
TF-IDF + BERT	0.62	0.62	0.62

Table 2: Macro-Averaged Precision, Recall, and F1-score for Different Models

Tamil	Malayalam
0.50	0.49

Table 3: Macro F1-scores of BERT embedding for predicted dataset

both individual models.

Test Data Predictions:

All three models made predictions on a separate test dataset after training. The predicted dataset had different results for the three models. The BERT-trained model had the most evenly distributed dataset. The hybrid model found very few abusive comments and labeled most of them as not abusive. In contrast, the TF-IDF model predominantly labels the entire test dataset for both Tamil and Malayalam as non-abusive. The performance of the proposed methodologies are shown in Tables 2 and 3.

6 Limitations

These languages are low-resource languages, hence we had lesser data points to train the model. We used fewer embeddings, so the model couldn't uncover all of the patterns required to classify the comments, which affected results. The approach for generating the feature representation was one of the reasons for obtaining low performance scores. Fine-tuning a pretrained model or using more data points to enable the model to uncover more hidden patterns could improve the model performance.

7 Conclusion

This paper presents the system description of the HTMS team for detecting abusive comments in Malayalam and Tamil against women. We experimented with three distinct methods for classifying abusive and non-abusive comments in Dravidian languages. We trained a model with fused embedding, which outperformed other models. After training and testing the datasets with these models, we observed that the BERT-only model provided excellent results. However, the TF-IDF model classified all comments as non-abusive, leading to highly inaccurate results.

References

Amena Akter Aporna, Istinub Azad, Nibraj Safwan Amlan, Md Humaion Kabir Mehedi, Mohammed Julfikar Ali Mahbub, and Annajiat Alim Rasel. 2022. Classifying offensive speech of bangla text and analysis using explainable ai. In *International Conference on Advances in Computing and Data Sciences*, pages 133–144. Springer.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Bagus Tri Yulianto Darmawan, Bassamtiano Renaufalgi Irnawan, and Yoshimi Suzuki. 2023. Indonesian hate speech and abusive tweets classification with deep learning pre-trained language models. In *2023 6th International Conference of Computer and Informatics Engineering (IC2IE)*, pages 30–35. IEEE.

Lennart Justen, Kilian Müller, Marco Niemann, and Jörg Becker. 2022. No time like the present: Effects of language change on automated comment moderation. In *2022 IEEE 24th Conference on Business Informatics (CBI)*, volume 1, pages 40–49. IEEE.

Sakshi Kalraa, Mehul Agrawala, and Yashvardhan Sharmaa. 2021. Detection of threat records by analyzing the tweets in urdu language exploring deep learning transformer-based models. In *Proc. CEUR Workshop*, pages 1–7.

KM Kavitha, Asha Shetty, Bryan Abreo, Adline D'Souza, and Akarsha Kondana. 2020. Analysis and classification of user comments on youtube videos. *Procedia Computer Science*, 177:593–598.

Shanmuga V Kogilavani, Subramanian Malliga, KR Jaiabinaya, M Malini, and M Manisha Kokila. 2023. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*, 81:630–633.

Tanjim Mahmud, Tahmina Akter, Mohammad Kamal Uddin, Mohammad Tarek Aziz, Mohammad Shahadat Hossain, and Karl Andersson. 2024. Machine learning techniques for identifying child abusive texts in online platforms. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Ari Muzakir, Kusworo Adi, and Retno Kusumaningrum. 2022. Classification of hate speech language detection on social media: Preliminary study for improvement. In *International Conference on Networking, Intelligent Systems and Security*, pages 146–156. Springer.

SN Prasanth, R Aswin Raj, P Adhithan, B Premjith, and Soman Kp. 2022. Cen-tamil@ dravidianlangtech-acl2022: Abusive comment detection in tamil using tf-idf and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 70–74.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022a. Overview of abusive comment detection in tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022b. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Dhanya Sathyan, Kalpathy Balakrishnan Anand, Aravind Jaya Prakash, and Bhavukam Premjith. 2018. Modeling the fresh and hardened stage properties of self-compacting concrete using random kitchen sink algorithm. *International journal of concrete structures and materials*, 12:1–10.

Malliga Subramanian, Kogilavani Shanmugavadivel, Nandhini Subbarayan, Adhithiya Ganesan, Deepti Ravi, Vasanth Palanikumar, and Bharathi Chakravarthi. 2023. On finetuning adapter-based transformer models for classifying abusive social media tamil comments.

Team_Catalysts@DravidianLangTech 2025: Leveraging Political Sentiment Analysis using Machine Learning Techniques for Classifying Tamil Tweets

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2], **Subhadevi K**[1], **Sowbharanika Janani J S**[1], **Rahul K**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{subhadevik,sowbharanikajananijs,rahulk}.22aid@kongu.edu

Abstract

This work proposed a methodology for assessing political sentiments in Tamil tweets using machine learning models. The approach addressed linguistic challenges in Tamil text, including cleaning, normalization, tokenization, and class imbalance, through a robust preprocessing pipeline. Various models, including Random Forest, Logistic Regression, and CatBoost, were applied, with Random Forest achieving a macro F1-score of 0.2933 and securing 8th rank among 153 participants in the Codalab competition. This accomplishment highlights the effectiveness of machine learning models in handling the complexities of multilingual, code-mixed, and unstructured data in Tamil political discourse. The study also emphasized the importance of tailored preprocessing techniques to improve model accuracy and performance. It demonstrated the potential of computational linguistics and machine learning in understanding political discourse in low-resource languages like Tamil, contributing to advancements in regional sentiment analysis.

Keywords: Sentiment Analysis, Machine Learning, Tamil Tweets, Political Sentiments, Random Forest, Class Imbalance, Natural Language Processing (NLP), Tokenization, Computational Linguistics, Multilingual Sentiment Analysis.

1 Introduction

Sentiment analysis, a key area of natural language processing (NLP), is vital for understanding public opinion, especially on social media platforms like Twitter. It helps monitor shifts in sentiment, offering insights into political discourse and its implications for governance.

This study focuses on sentiment analysis of Tamil tweets, presenting unique challenges due to Tamil's complex script, informal idioms, and limited annotated datasets. Traditional techniques, designed for languages like English, often fall short, highlighting the need for language-specific approaches.

Previous studies, such as Mutanov et al. (2021), addressed imbalanced sentiment classes using resampling techniques, with logistic regression, decision trees, and random forests showing strong performance. Liu et al. (2017) highlighted the effectiveness of SVM in sentiment classification, emphasizing feature selection, while Bouazizi and Ohtsuki (2018) introduced "quantification" to capture multiple sentiments within a single post.

This work employs machine learning models like Random Forest, Logistic Regression, Naive Bayes, Decision Tree, AdaBoost, Gradient Boost, and CatBoost to classify Tamil tweets as positive, negative, or neutral. Techniques such as text normalization, stop-word removal, and tokenization with Stanza's Tamil NLP model were applied, contributing to the advancement of multilingual sentiment analysis, particularly for regional languages like Tamil.

2 Related Works

Political sentiment analysis became an important tool for assessing public opinion, particularly in political contexts. Elghazaly et al. (2016) employed SVM and Naïve Bayes to classify Twitter data during Egypt's 2012 presidential election. The study used Term Frequency-Inverse Document Frequency (TF-IDF) for vectorization and found that Naïve Bayes outperformed SVM in terms of accuracy and error rates. Similarly, Bose et al. (2019) utilized sentiment analysis with the NRC Emotion Lexicon and ParallelDots AI APIs to monitor the 2017 Gujarat Legislative Assembly Election. Their methodology categorized tweets as positive, negative, or neutral, effectively summarizing public sentiment. In contrast, Singhal et al. (2015) focused on a context-aware, semantics-based approach to predict election results by analyzing Twitter data from

the 2019 Indian General Election. They proposed a rules-based system to extract sentiment, which aligned with actual election outcomes, highlighting the importance of domain-specific techniques in political sentiment analysis.

More advanced techniques, such as Long Short Term Memory (LSTM) networks, were used to improve sentiment classification accuracy. For example, Ansari et al. (2020) employed LSTM to analyze Twitter data from the 2019 Indian General Elections. They evaluated the performance of LSTM against classical machine learning models and found that deep learning models, particularly LSTM, outperformed traditional methods. Similarly, Pinto and Murari (2019) investigated the use of LSTM for real-time political sentiment analysis by evaluating tweets about the Ayodhya dispute. Their findings demonstrated that LSTM effectively tackled challenges presented by multiple languages and large datasets. Desai and Mehta (2016) compared various sentiment analysis algorithms, including Naïve Bayes, SVM, and neural networks. Their survey highlighted the benefits of deep learning models and demonstrated their efficiency in categorizing unstructured Twitter data as positive, negative, or neutral.

In addition to these strategies, hybrid approaches combining lexicon-based and machine learning techniques were shown to improve sentiment classification accuracy. For example, Ringsquandl and Petkovic (2013) proposed a hybrid strategy for improving aspect extraction by combining noun phrase frequency and Pointwise Mutual Information (PMI), resulting in higher sentiment classification accuracy. Similarly, Thavareesan and Mahesan (2019) tested numerous sentiment analysis techniques on Tamil literature, including lexicon-based, machine learning, and hybrid methods. Their research revealed that the supervised machine learning methodology, which used fastText and customized corpora, outperformed other methods, achieving 0.79% accuracy. Furthermore, Anish and Sumathy (2022) focused on Tamil political evaluations, employing an SVM approach to address language-specific problems such as noise and sarcasm and proposed improvements through context-based sentiment extraction.

Traditional machine learning methods, such as Naïve Bayes and SVM, offered quick and interpretable solutions, while deep learning methods, such as LSTM, provided higher accuracy. The evolution of political sentiment analysis from traditional to deep learning and hybrid approaches reflected the growing sophistication in the field. Each methodology had its strengths, depending on the dataset's complexity, language, and sentiment classification granularity.

3 Problem and System Description

The system was designed to perform sentiment analysis on Tamil political tweets, which presented challenges due to their brief and unstructured nature, mixed-language usage, and the complexity of Tamil. Tweets often used code-switching between Tamil and English, making sentiment analysis more difficult. The goal was to effectively characterize political sentiments in Tamil tweets, addressing challenges such as data imbalance and the unique features of Tamil political discourse.

Sentiment analysis of Tamil political tweets was critical for determining public opinion on political issues. Social media platforms like Twitter (X) served as important sources of real-time political conversation, but research in this area, particularly for Tamil, was limited. This approach aimed to bridge gaps in multilingual sentiment analysis, especially for Dravidian languages.

The system employed a structured pipeline that began with data gathering from Tamil political tweets, followed by pre-processing to address language-specific issues. Stanza was utilized for tokenization, and TF-IDF was used to extract features and identify relevant words. The pre-processed dataset was then used to train various machine learning models for sentiment classification.

This study leveraged the dataset introduced in Ravikiran et al. (2022) on Dravidian Code-Mixed Offensive Span Identification, which is crucial for addressing language-specific challenges in sentiment analysis. Their dataset has significantly contributed to advancing the field of Dravidian language processing Chakravarthi et al. (2025).

3.1 Dataset Description

The dataset consists of Tamil political tweets with content and labels columns, and the labels reflect one of seven sentiment categories as shown in Table 1 below.

The dataset is divided into training, validation, and testing sets. Table 2 illustrates the distribution of the data.

Table 1: Labels

Categories
Substantiated
Sarcastic
Opinionated
Positive
Negative
Neutral
None of the above

Table 2: Dataset Description

Dataset	No. of Tweets
Train	4352
Validation	544
Test	544

4 Methodology

The following methodology describes the stages required in sentiment analysis of political tweets using several machine learning models. The approach consists of three major components: diagrammatic representation, preprocessing stages, and test data predictions.

4.1 Diagrammatic Representation of Proposed Work

The figure 1 below depicts the full sentiment analysis procedure. The process starts with data collection and progresses through preprocessing, dataset balance, feature extraction, model training, evaluation, and prediction on test data. This end-to-end procedure guarantees that raw text data is handled, models are properly trained, and predictions are produced on previously unknown data.

4.2 Preprocessing Steps

This stage converts raw textual data into a format appropriate for model training. The first step in preprocessing is text cleaning. The text is normalized by deleting unnecessary Unicode characters, maintaining only the required script, and removing special characters and numerals. Commonly used spoken versions are normalized (for example, similar-sounding words are replaced), and specific characters are mapped to their basic forms.

Tokenization is then performed using the proper tokenization library. This transforms each sentence into a list of tokens for further processing. Class balancing is also used, which involves upsampling

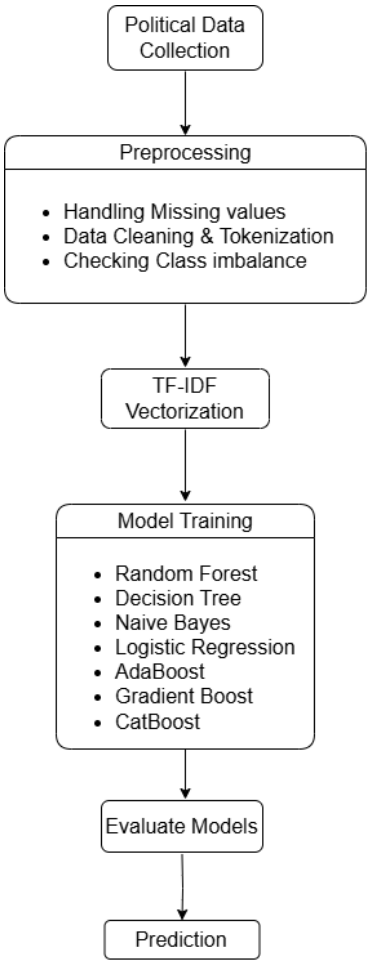


Figure 1: Proposed system pipeline

underrepresented classes to ensure equal representation of all sentiment categories.

4.3 Predictions on Test Data

After the models have been trained, predictions are made using the test dataset. To guarantee consistency, the test data goes through the same preparatory stages (normalization, text cleaning, and tokenization). The TF-IDF vectorizer that was fitted to the training data is utilized to convert the test data into numerical feature vectors.

The trained models are then used to estimate the sentiment of the test data. For example, the Random Forest model is used to assign tweets to one of seven attitude categories: substantive, sarcastic, opinionated, positive, negative, neutral, or none of the above. Once predictions are created, they are saved in a separate column of the test dataset for further examination.

5 Results

The machine learning models performance is measured using measures such as accuracy, precision, recall, and F1-score. During training and testing, each model demonstrated the following accuracies, as shown in Table 3. The code used for preprocessing and analysis can be found in this GitHub repository: Political Multiclass Sentiment Analysis

Table 3: Tarin and Test Accuracy

Model	Train	Test
Random Forest	0.94	0.84
Decision Tree	0.94	0.83
CatBoost	0.70	0.61
Gradient Boost	0.68	0.58
Logistic Regression	0.62	0.52
Naive Bayes	0.57	0.50
AdaBoost	0.28	0.29

To evaluate the performance of various machine learning models in classifying attitudes in political tweets, we analyzed them using key metrics such as accuracy, precision, recall, and F1-score. The results from different models were compared and visualized in the figures below.

Figure 2 presented a comparative analysis of the accuracy of multiple classification algorithms, including Random Forest, Decision Tree, CatBoost, Gradient Boost, and Logistic Regression. The Random Forest model achieved the highest accuracy (0.84), followed closely by the Decision Tree (0.83). CatBoost, Gradient Boost, and Logistic Regression demonstrated lower accuracy levels, indicating their limited effectiveness for this task.

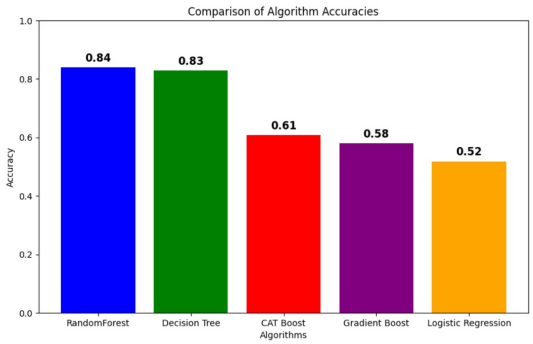


Figure 2: Comparision of Algorithm Accuracies

Figure 3 provided a detailed classification report for the best-performing model, Random Forest, with an overall test accuracy of 83.99%. The model's precision, recall, and F1-score were reported for each sentiment category, demonstrating strong performance in distinguishing between various political attitudes. The confusion matrix further highlighted the model's effectiveness in correctly classifying instances while showing potential misclassifications across specific categories.

```
Test Data Evaluation:
Accuracy: 0.8399790136411333
Classification Report:
                    precision    recall  f1-score   support

          Negative       0.90      0.87      0.89       272
           Neutral       0.85      0.76      0.80       272
None of the above       0.91      0.98      0.94       272
       Opinionated       0.73      0.71      0.72       273
          Positive       0.85      0.87      0.86       272
         Sarcastic       0.84      0.83      0.84       272
     Substantiated       0.80      0.86      0.83       273

          accuracy                           0.84      1906
         macro avg       0.84      0.84      0.84      1906
      weighted avg       0.84      0.84      0.84      1906

Confusion Matrix:
 [[237   3   4   7   2   8  11]
 [  5 208   2  18  14   7  18]
 [  0   0 267   0   0   2   3]
 [ 11  15   6 193  18  21   9]
 [  4   6   6  11 236   1   8]
 [  2   8   5  18   4 226   9]
 [  3   6   4  19   4   3 234]]
```

Figure 3: Classification Report and Confusion Matrix

These results indicated that ensemble models like Random Forest and Decision Tree outperformed other approaches in classifying political sentiments. Their high accuracy and balanced precision-recall scores made them suitable choices for this task. However, further improvements could have been made by optimizing hyperparameters or incorporating advanced deep learning techniques.

6 Conclusion

This study focused on classifying political emotions in Tamil tweets using machine learning models. Ensemble methods like Random Forest and Decision Tree achieved high accuracy by capturing complex patterns, while CatBoost and Gradient Boosting showed promising results. Simpler models like Logistic Regression and Naive Bayes struggled with data complexity, and AdaBoost highlighted the need for more robust models. These findings demonstrate the effectiveness of ensemble techniques in handling linguistic nuances and data imbalances. The study underscores the importance of advanced Machine Learning techniques in political sentiment analysis and suggests exploring deep learning or hybrid models for improved accuracy and deeper insights.

References

Mohd Zeeshan Ansari, Mohd-Bilal Aziz, MO Siddiqui, H Mehra, and KP Singh. 2020. Analysis of political sentiment orientations on twitter. *Procedia computer science*, 167:1821–1828.

Rajesh Bose, Raktim Kumar Dey, Sandip Roy, and Debabrata Sarddar. 2019. Analyzing political sentiment using twitter data. In *Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 2*, pages 427–436. Springer.

Mondher Bouazizi and Tomoaki Ohtsuki. 2018. Multi-class sentiment analysis in twitter: What if classification is not the answer. *IEEE access*, 6:64486–64502.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Mitali Desai and Mayuri A Mehta. 2016. Techniques for sentiment analysis of twitter data: A comprehensive survey. In *2016 international conference on computing, communication and automation (ICCCA)*, pages 149–154. IEEE.

Tarek Elghazaly, Amal Mahmoud, and Hesham A Hefny. 2016. Political sentiment analysis using twitter data. In *Proceedings of the International Conference on Internet of things and Cloud Computing*, pages 1–5.

Yang Liu, Jian-Wu Bi, and Zhi-Ping Fan. 2017. Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. *Expert Systems with Applications*, 80:323–339.

Galimkair Mutanov, Vladislav Karyukin, and Zhanl Mamykova. 2021. Multi-class sentiment analysis of social media data with machine learning algorithms. *Computers, Materials & Continua*, 69(1).

Joylin Priya Pinto and Vijaya Murari. 2019. Real time sentiment analysis of political twitter data using machine learning approach. *International Research Journal of Engineering and Technology (IRJET)*, 6(4):4124–4129.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on Toxic Span Identification in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Martin Ringsquandl and Dusan Petkovic. 2013. Analyzing political sentiment on twitter. In *2013 AAAI Spring Symposium Series*.

Kartik Singhal, Basant Agrawal, and Namita Mittal. 2015. Modeling indian general elections: sentiment analysis of political twitter data. In *Information Systems Design and Intelligent Applications: Proceedings of Second International Conference INDIA 2015, Volume 1*, pages 469–477. Springer.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on industrial and information systems (ICIIS)*, pages 320–325. IEEE.

InnovationEngineers@DravidianLangTech 2025: Enhanced CNN Models for Detecting Misogyny in Tamil Memes Using Image and Text Classification

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2], **Pooja Sree M**[1], **Palanimurugan V**[1],**Roshini Priya K** [1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{poojasreem,palanimuruganv,roshinipriyak}.22aid@kongu.edu

Abstract

The rise of misogynistic memes on social media posed challenges to civil discourse. This paper aimed to detect misogyny in Dravidian language memes using a multimodal deep learning approach. We integrated Bidirectional Encoder Representations from Transformers (BERT), Long Short-Term Memory (LSTM), EfficientNet, and a Vision Language Model (VLM) to analyze textual and visual information. EfficientNet extracted image features, LSTM captured sequential text patterns, and BERT learned language-specific embeddings. Among these, VLM achieved the highest accuracy of 85.0% and an F1-score of 70.8, effectively capturing visual-textual relationships. Validated on a curated dataset, our method outperformed baselines in precision, recall, and F1-score. Our approach ranked 12th out of 118 participants for the Tamil language, highlighting its competitive performance. This research emphasizes the importance of multimodal models in detecting harmful content. Future work can explore improved feature fusion techniques to enhance classification accuracy.

Keywords: LSTM, BERT, EfficientNet, Vision Language Model, Meme Classification

1 Introduction

Social media platforms have developed into places for entertainment and idea sharing, but they have also made it possible for harmful information, such as misogynistic memes, to proliferate. These text-and-image memes frequently spread harmful viewpoints and disrupt online conversation. In Dravidian languages, which are spoken in southern India and nearby areas and are frequently underrepresented in language processing studies, it is particularly difficult to detect such material. By creating a binary classification model that classifies information as either abusive or non abusive, this work aims to detect misogyny in memes across Dravidian languages. Using a multimodal deep learning architecture, our method analyzes both the text and images in memes by combining multiple powerful models: Vision-Language Models (VLM) to comprehend the relationship between images and text, EfficientNet to extract image features efficiently, BERT to process language-specific text embeddings, and LSTM to capture the flow of text sequences. The VLM achieves the highest accuracy among these models, highlighting the significance of analyzing both visual and textual information simultaneously. Our model performs well on a curated dataset, surpassing previous methods and helping to combat abusive content in low-resource languages.

2 Literature Survey

The growing prevalence of online hate speech, particularly in the form of misogyny and trolling, has led to a surge in research focusing on the detection of such harmful content. The detection of misogyny in online content, particularly through memes, has been extensively studied in recent years.

Ponnusamy et al. (2024) introduced an annotated dataset for misogyny detection in Tamil and Malayalam memes. The study highlights challenges in identifying offensive content in low-resource languages due to cultural and contextual nuances. The dataset, sourced from social media, is manually labeled and evaluated using NLP techniques. Their work emphasizes the importance of multimodal approaches, considering both textual and visual elements in memes. This research aids in developing AI systems to detect online misogyny and improve fairness in content moderation.

Jindal et al. (2024) introduced MISTRA, a novel approach that combines text and image features for misogyny detection, emphasizing the effectiveness of fusion models to detect subtle forms of misogyny in multimodal platforms like memes and images. Chinivar et al. (2024) proposed V-LTCS, focusing

on the importance of selecting appropriate backbone networks for multimodal misogynous meme detection, which significantly impacts the performance of detecting misogynistic content in memes. Raja et al. (2023) applied a transfer learning approach with adaptive fine-tuning for fake news detection in Dravidian languages, offering a methodology that could be valuable for language-specific challenges in misogyny detection within multilingual contexts. Kumari et al. (2024) introduced M3Hop-CoT, employing a multimodal, multi-hop chain-of-thought process for meme identification, which emphasizes the contextual relationships between visual and textual elements for improved misogyny detection.

Srivastava (2024) proposed an early fusion model with graph networks for meme detection, demonstrating the advantages of combining multimodal features before processing to capture complex patterns in memes.Rizzi et al. (2023) discussed biases in misogyny detection models and stress the importance of addressing these biases for fairer and more equitable detection systems. Chakravarthi et al. (2024) provided a comprehensive evaluation of multitask meme classification, focusing on detecting both misogynistic and troll content in memes, aiming to improve classification accuracy and address various forms of online abuse.

Anzovino et al. (2018) explored automatic identification and classification of misogynistic language on Twitter, offering early insights into the challenge of detecting misogynistic content in text-based platforms. Plaza-Del-Arco et al. (2020) extended this work by focusing on the detection of misogyny and xenophobia in Spanish tweets, contributing to multilingual approaches in hate speech and misogyny detection. Frenda et al. (2019) investigated online hate speech against women, highlighting challenges in identifying misogyny and sexism on platforms like Twitter. Chakravarthi et al. (2025) presented the findings of the misogyny meme detection task in Dravidian languages at NAACL 2025, analyzing various machine learning and deep learning models. The study highlights dataset creation, annotation challenges, and the importance of multimodal approaches for effective detection.

Kiela et al. (2020) introduced the Hateful Memes Challenge, emphasizing the detection of hate speech in multimodal memes, which is a pivotal work in the multimodal analysis of harmful online content. Zhu (2020) enhanced multimodal transformer models for the Hateful Meme Challenge, showcasing how external labels and pretraining can improve meme classification accuracy. Zia et al. (2021) expanded on this by classifying memes beyond hate speech, tackling the challenge of detecting misogynistic and sexist memes specifically. Aloysius and Tamil Selvan (2023) addressed the reduction of false negatives in multi-class sentiment analysis, which is an important consideration for improving the accuracy of automated sentiment detection models, including those used for identifying misogynistic content.

3 Dataset Description

The dataset is made up of a series of pictures and a matching CSV file with three essential elements as given image id, labels and transcriptions. Each image in the folder is uniquely identified by its image id, which makes sure that every image is easily identifiable. The labels field indicates the type of content by classifying each image as either non-misogynistic or misogynistic. The textual information linked to each image is included in the transcriptions field, which gives the visual data context.

4 Methodology

Our model was developed as part of the Dravidian-LangTech 2025 Shared Task for detecting misogyny in Tamil memes. The task challenged participants to create robust multimodal models capable of processing both text and visual elements.

4.1 Dataset Preprocessing

Dataset preprocessing involved handling missing values, encoding labels, and splitting the dataset into training (1,136 memes), validation (284), and testing (356), totaling 1,776 memes. Text preprocessing included lowercasing, removing special characters, stopwords, and padding sequences. Image preprocessing involved resizing, standardizing pixel values, and data augmentation. These steps ensure standardized, noise-free input for optimal model performance. The dataset split is given below in Table 1.

4.1.1 Text Preprocessing

In order to prevent case sensitivity, all text was converted to lowercase, cleaning and standardizing the text data for model input. Only relevant Tamil and English characters were left after special characters, emojis, URLs, and other unnecessary

Dataset Split	Number of Memes
Training Set	1,136
Validation Set	284
Test Set	356
Total	1,776

Table 1: Dataset Description

symbols were eliminated using regular expressions. They removed stopwords such as "enna" (what) and "ipo" (now), and they padded or cleaned sequences to a specified length of 100. To expand the vocabulary and maintain uniformity throughout the dataset, words that appeared less than five times were substituted with placeholder tokens.

4.1.2 Image Preprocessing

Image data was standardized to ensure compatibility with the deep learning model. First, all images were resized to a target size of 224x224 pixels, which is the required input size for the EfficientNetB0 model. To enhance training efficiency, pixel values were normalized to a range of [0, 1] by dividing by 255, ensuring faster convergence during model training. Additionally, images were processed in batches, converting each image into an array and stacking them into a single dataset for efficient handling during training. These preprocessing steps ensured that the image data was optimized for input into the model.

4.2 Models

4.2.1 Vision Language Model

A Vision-Language Model (VLM), which achieves 85% accuracy and a 70.8 F1-score, successfully combines text and images. Although accuracy indicates general correctness, efficiency could be improved by controlling class imbalance and increasing precision. The classification report is shown in Figure 1. VLMs excel in tasks like image captioning, visual quality assurance, and image-text matching, leveraging CNNs for images and Transformer-based models for text.

4.2.2 EfficientNet

EfficientNet optimizes accuracy and efficiency using compound scaling for tasks like object detection and image classification. With 82% accuracy and a 72.5 F1 score, it balances precision and recall well.The classification report is shown in Figure 2. Further tuning and data augmentation could enhance performance while maintaining reliability in

Classification Report:

	precision	recall	f1-score	support
0	0.88	0.92	0.90	210
1	0.70	0.60	0.70	74
accuracy			0.85	284
macro avg	0.79	0.76	0.78	284
weighted avg	0.84	0.85	0.83	284

Figure 1: Classification Report for VLM

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.90	0.88	210
1	0.68	0.65	0.72	74
accuracy			0.82	284
macro avg	0.78	0.78	0.73	284
weighted avg	0.82	0.82	0.81	284

Figure 2: Classification Report for EfficientNet

reducing false predictions.

4.2.3 BERT

BERT enhances semantic understanding for NLP tasks using a bidirectional transformer. With 79.5% accuracy and a 65 F1 score, it performs well but struggles with class imbalance. Techniques like oversampling, class weighting, domain-specific models, hyperparameter tuning, and data augmentation can improve performance.

4.2.4 LSTM

LSTM, a type of RNN, excels in sequential tasks like NLP and time series forecasting by retaining long-term dependencies. With 75% accuracy and a 67.5 F1 score, it performs well but can improve on imbalanced data. Enhancements like more layers, bidirectional LSTMs, fine-tuning, class balancing, pre-trained embeddings, or attention mechanisms can boost performance.

5 Workflow

The workflow begins with text and image preprocessing, where BERT extracts textual features and CNN captures image features. These features are fused and fed into a training model comprising LSTM, EfficientNet, and VLM. LSTM processes sequential text patterns, EfficientNet refines image features, and VLM models text-image interactions. After training, the model is evaluated for accuracy,

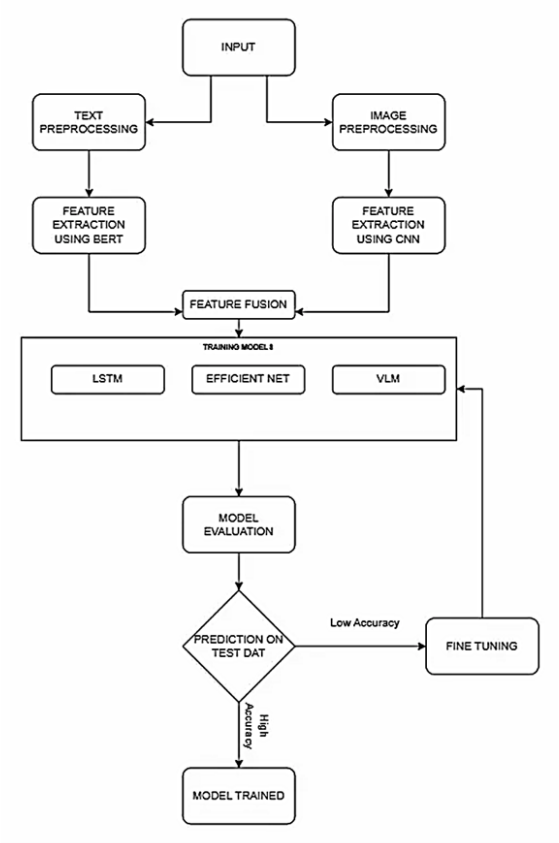


Figure 3: Proposed System Workflow

precision, recall, and F1-score. If accuracy is high, the model is finalized. Otherwise, fine-tuning is performed to enhance performance. Figure 3 illustrates this iterative process for effective misogyny detection in Dravidian language memes.

6 Result and Discussion

This study investigated the performance of many deep learning architectures for multimodal classification tasks, such as Vision-Language Models (VLM), BERT, LSTM, and EfficientNet. The comparison analysis demonstrates how various feature extraction and fusion techniques affect attaining peak performance.And our model ranked 12th out of 118 participants in the Tamil language track, showcasing its competitive performance in multimodal misogyny detection.The model accuracy performance comparision has been shown in Table 2.

VLM achieved the highest accuracy (85.0%), demonstrating its effectiveness in integrating textual and visual features. However, EfficientNet outperformed other models in F1-score (72.5%), indicating a better balance between precision and recall. While BERT excelled in text-based tasks, it lagged behind multimodal approaches. LSTM, though less accurate overall, maintained a competitive F1-score (67.5%), making it suitable for sequential feature extraction.The preprocessing of images and implementation details can be found in our GitHub repository InnovationEngineers Misogyny meme detection.

Models	Accuracy	F1-Score
VLM	85.0	70.8
EfficientNet	82.0	72.5
BERT	79.5	65.0
LSTM	75.0	67.5

Table 2: Model Performance Comparison

7 Conclusion

This study emphasized how Vision-Language Models (VLM) are a great help for multimodal categorization problems. VLM performs better than other models, exhibiting a superior capacity to integrate both textual and visual data with an accuracy of 85%. A better balance between precision and recall is demonstrated by EfficientNet's superior F1-score (72.5%), but VLM's total performance emphasizes

the value of multimodal learning in improving classification accuracy. In sequential and text-based feature extraction tasks, BERT and LSTM offer useful insights, but multimodal techniques like VLM and EfficientNet outperform them.

References

C. Aloysius and P. Tamil Selvan. 2023. Reduction of false negatives in multi-class sentiment analysis. *Bulletin of Electrical Engineering and Informatics*, 12(2):1209–1218.

M. Anzovino, E. Fersini, and P. Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

B. R. Chakravarthi, S. Rajiakodi, R. Ponnusamy, K. Pannerselvam, A. K. Madasamy, R. Rajalakshmi, H. Ramakrishna Iyer Lekshmi Ammal, A. Kizhakkeparambil, S. S. Kumar, B. Sivagnanam, and C. Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the LT-EDI@EACL 2024*.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

S. Chinivar, M. S. R., J. S. A., and K. R. V. 2024. V-ltcs: Backbone exploration for multimodal misogynous meme detection. *Natural Language Processing*, 100109.

S. Frenda, B. Ghanem, M. Montes y Gómez, and P. Rosso. 2019. Online hate speech against women: automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

N. Jindal, P. K. Kumaresan, R. Ponnusamy, S. Thavareesan, S. Rajiakodi, and B. R. Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing*, 100073.

D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

G. Kumari, K. Jain, and A. Ekbal. 2024. M3hop-cot: Misogynous meme identification with multimodal multi-hop chain-of-thought. In *Proceedings of the SemEval-2022 Task 5 (MAMI Task)*.

F.-M. Plaza-Del-Arco, M.D. Molina-González, L.A. Ureña-López, and M.T. Martín-Valdivia. 2020. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–19.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

E. Raja, B. Soni, and S. K. Borgohain. 2023. Fake news detection in dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 106877.

G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, and E. Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, 60(6):103474.

H. Srivastava. 2024. Misogynistic meme detection using early fusion model with graph network. In *Proceedings of the SemEval-2022 Task 5*.

R. Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv:2012.08290 [cs.CL]*.

H. B. Zia, I. Castro, and G. Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the Workshop on Online Abuse and Harms (WOAH 2021)*, pages 1–10.

MysticCIOL@DravidianLangTech 2025: A Hybrid Framework for Sentiment Analysis in Tamil and Tulu Using Fine-Tuned SBERT Embeddings and Custom MLP Architectures

Minhaz Chowdhury, Arnab Laskar, Taj Ahmad, Azmine Toushik Wasi[†]
Shahjalal University of Science and Technology, Sylhet, Bangladesh
[†]**Correspondence:** azmine32@student.sust.edu

Abstract

Sentiment analysis is a crucial NLP task used to analyze opinions in various domains, including marketing, politics, and social media. While transformer-based models like BERT and SBERT have significantly improved sentiment classification, their effectiveness in low-resource languages remains limited. Tamil and Tulu, despite their widespread use, suffer from data scarcity, dialectal variations, and code-mixing challenges, making sentiment analysis difficult. Existing methods rely on traditional classifiers or word embeddings, which struggle to generalize in these settings. To address this, we propose a hybrid framework that integrates fine-tuned SBERT embeddings with a Multi-Layer Perceptron (MLP) classifier, enhancing contextual representation and classification robustness. Our framework achieves validation F1-scores of 0.4218 for Tamil and 0.3935 for Tulu and test F1-scores of 0.4299 in Tamil and 0.1546 on Tulu, demonstrating its effectiveness. This research provides a scalable solution for sentiment classification in low-resource languages, with future improvements planned through data augmentation and transfer learning. Our full experimental codebase is publicly available at: github.com/ciol-researchlab/NAACL25-Mystic-Tamil-Sentiment-Analysis.

1 Introduction

Sentiment analysis, or opinion mining, is a critical task in Natural Language Processing (NLP) that involves determining the sentiment or emotional tone expressed in a given text. It plays a vital role in analyzing customer feedback, public opinion, and social media discourse, influencing industries such as marketing, politics, and e-commerce (Sebastiani, 2002). Traditional sentiment analysis methods relied on lexicon-based or machine learning approaches, but recent advancements in deep learning and transformer-based architectures have significantly improved performance. Models like BERT and SBERT can capture complex contextual relationships, making sentiment classification more accurate (Devlin, 2018; Reimers, 2019). However, while these models excel in high-resource languages like English, their effectiveness in low-resource languages remains limited. Sentiment analysis in underrepresented languages, such as Tamil and Tulu, presents unique challenges, including morphological richness, dialectal variations, and a lack of high-quality annotated datasets.

Sentiment analysis in underrepresented languages like Tamil and Tulu faces challenges such as morphological complexity, dialectal variations, and data scarcity. Despite being spoken by millions, these languages lack computational tools and annotated datasets (Joshi et al., 2020). Traditional classifiers and word embeddings struggle in resource-constrained settings (Hedderich et al., 2020). Additionally, code-switching and syntactic variations complicate sentiment classification, as monolingual models fail to capture linguistic nuances (Hegde et al., 2023b). Addressing these challenges requires adaptable frameworks that integrate advanced NLP techniques while mitigating data limitations.

To tackle these challenges, this study proposes a hybrid framework that integrates fine-tuned SBERT embeddings with a custom Multi-Layer Perceptron (MLP) classifier. The SBERT model generates high-dimensional contextual embeddings, capturing intricate linguistic patterns in Tamil and Tulu text. These embeddings are then processed using an optimized MLP classifier incorporating dropout regularization and ReLU activation to enhance robustness and prevent overfitting (Hwang and Jeong, 2023). The proposed framework achieves validation F1-scores of 0.4218 for Tamil and 0.3935 for Tulu, demonstrating its effectiveness in sentiment classification for low-resource languages. This research contributes to the broader field of NLP by

establishing a scalable and adaptable classification pipeline. Future work will explore data augmentation, transfer learning, and hyperparameter tuning to further optimize performance, ultimately fostering more inclusive AI solutions for underrepresented linguistic communities (Feng et al., 2021).

2 Problem Description

Problem Statement. The 7th task, Sentiment Analysis in Tamil and Tulu, in the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025) at NAACL 2025 (Durairaj et al., 2025; Chakravarthi et al., 2020; Hegde et al., 2022, 2023a; S. K. et al., 2024) focuses on text classification in low-resource settings. Tamil and Tulu, despite their rich linguistic heritage, face significant challenges due to the scarcity of labeled datasets and the complexity of their syntactic and morphological structures. Traditional sentiment classification models struggle in these languages, particularly in multilingual and code-mixed contexts, where words from different languages are interwoven. Key challenges in low-resource text classification include limited high-quality annotated datasets, complicating supervised learning, and issues like code-mixing and dialectal variations, which introduce inconsistencies in syntax, morphology, and orthography. Small dataset sizes increase overfitting risks, reducing generalization. Additionally, pre-trained embeddings for these languages are often unavailable or underdeveloped, limiting the effectiveness of transformer-based models. Overcoming these challenges requires innovative approaches that utilize contextual embeddings while addressing data scarcity issues.

Dataset. This study utilizes Tamil and Tulu sentiment analysis datasets, split into training, validation, and test sets (Durairaj et al., 2025; Chakravarthi et al., 2020; Hegde et al., 2022, 2023a; S. K. et al., 2024). The datasets contain labeled textual inputs across various sentiment categories, helping to train and evaluate classification models effectively. Tamil has 31,122 training, 3,843 validation, and 3,459 test samples, while Tulu consists of 13,308 training, 1,643 validation, and 1,479 test samples. The test sets lack sentiment labels, requiring model predictions for evaluation. These datasets capture diverse linguistic structures, including code-mixed text, posing challenges for sentiment classification. A summary of dataset statistics is presented in Table 1.

Table 1: Overview of Tamil and Tulu datasets.

Dataset	Language	Entries
Training Set	Tamil	31,122
Validation Set	Tamil	3,843
Test Set	Tamil	3,459
Training Set	Tulu	13,308
Validation Set	Tulu	1,643
Test Set	Tulu	1,479

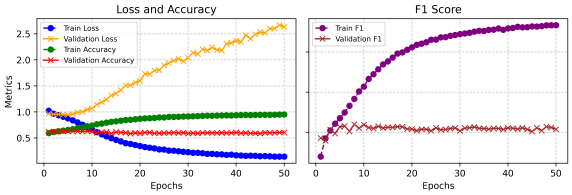


Figure 1: Training and validation metrics for Tamil language.

3 Methodology

We employed a comprehensive methodology to classify textual data effectively using an advanced hybrid pipeline that combines Sentence-BERT (SBERT) embeddings with a custom Multi-Layer Perceptron (MLP) architecture. Our approach leverages SBERT to generate rich, contextualized representations of Tamil and Tulu text, capturing semantic nuances often missed by traditional embeddings. These embeddings are then processed through an optimized MLP classifier, incorporating dropout regularization and ReLU activation to enhance robustness and prevent overfitting. This hybrid framework enables efficient sentiment classification in low-resource settings, addressing key challenges such as code-mixing and dialectal variations.

Data Preprocessing. We mapped textual sentiment labels to numerical values to ensure compatibility with machine learning models. For Tamil, we utilized the pre-trained SBERT model *l3cube-pune/indic-sentence-similarity-sbert* (Deode et al., 2023), while for Tulu, we employed *m3hrdadfi/zabanshenas-roberta-base-mix* (Farahani, 2021). These models are well-suited for capturing semantic similarities, making them effective for sentiment classification. Additionally, all preprocessing operations were conducted in a GPU-enabled environment to optimize efficiency and reduce computational overhead.

Embedding Generation. To generate embeddings,

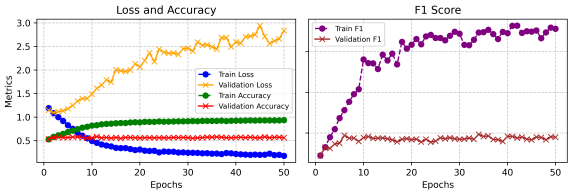


Figure 2: Training and validation metrics for Tulu language.

we tokenized the textual data using the specialized tokenizer from each respective SBERT model, ensuring proper padding and truncation to a maximum sequence length of 1024 tokens (Reimers, 2019). We extracted embeddings from the [CLS] token in the last hidden layer, as it provides a high-dimensional semantic representation of the input text (Devlin, 2018). These embeddings were then stored as .pt files to avoid redundant computations. For training and validation, the embeddings were paired with numerical labels, while test set embeddings were stored separately for later evaluation.

MLP Model Design. We designed two MLP models—one for Tamil and one for Tulu—to classify the high-dimensional embeddings effectively. Each MLP consisted of three fully connected layers with dimensions of 1024, 512, and the number of output classes, respectively. To enhance non-linearity, we applied ReLU activation functions after each hidden layer. Additionally, dropout layers with a probability of 0.3 were incorporated to mitigate overfitting (Srivastava et al., 2014). The final classification layer used a softmax activation function to generate class probabilities. For optimization, we utilized the Adam optimizer with a learning rate of 0.001 and CrossEntropyLoss as the loss function, ensuring stable convergence during training.

Training Process. We trained our model for 50 epochs, monitoring performance on the validation dataset after each epoch. To optimize memory usage and improve generalization, we shuffled the data and processed it in batches of 32. The model adjusted its weights using backpropagation, minimizing the loss function. We evaluated performance using accuracy, precision, recall, and F1-score (Powers, 2020), ensuring a comprehensive assessment of generalization capability. The model with the highest F1 score on the validation set was saved to disk, guaranteeing that only the most optimal version was used for final predictions. To maintain consistency across runs, we applied a fixed random seed.

Testing and Submission. For the test dataset, predictions were generated using the saved best model. The test embeddings were passed through the trained MLP, and class probabilities were computed. The final predictions were determined by selecting the class with the highest probability. A submission file containing the original text and its predicted label was prepared and saved in CSV format. This methodology ensured a systematic approach to embedding generation, model training, and evaluation, leveraging SBERT's semantic power and MLP's flexibility to achieve robust classification results (Deode et al., 2023; Farahani, 2021; Devlin, 2018).

4 Results and Discussion

Tamil Language Results. For the Tamil test dataset, predictions were generated using the best-performing model. The test embeddings were processed through the trained MLP, and class probabilities were computed. The final predictions were determined by selecting the class with the highest probability. A structured submission file was created, containing the original text and its predicted sentiment label, ensuring systematic evaluation. This methodology leveraged SBERT's semantic representation capabilities and MLP's adaptability to achieve robust classification results (Deode et al., 2023; Farahani, 2021; Devlin, 2018). During training, the Tamil language model exhibited consistent improvements in classification performance. By epoch 30, the training F1-score reached 0.8890, while the validation F1-score stabilized at 0.4218. Training beyond epoch 40 resulted in diminishing returns, with validation performance plateauing, suggesting that the model had effectively captured core linguistic patterns in the dataset. Precision and recall metrics showed balanced growth, with precision steadily increasing, indicating improved classification accuracy. However, recall demonstrated slight declines in later epochs, reflecting the challenge of maintaining generalization as the model became more confident in its predictions. Beyond epoch 40, the gap between training and validation performance widened, indicating overfitting tendencies (Ying, 2019). The training accuracy reached 93.22% by epoch 50, whereas validation accuracy remained at 56.42%, with a final validation F1-score of 0.3704. The validation loss peaked at 2.7500 by epoch 44, further confirming signs of overfitting.

Tulu Language Results. The Tulu language model followed a similar trend to the Tamil model, showing early improvements before reaching a plateau. Initial training began with a modest accuracy of 52.89% and an F1-score of 0.2897 at epoch 1. At this stage, validation accuracy and F1-score were recorded as 53.26% and 0.2879, respectively. By epoch 6, the model exhibited significant progress, achieving a training accuracy of 71.45% and an F1-score of 0.5126, while validation accuracy and F1-score reached 57.88% and 0.3887. The model's highest recorded validation F1-score of 0.3935 was observed at epoch 34 (Figure 2). However, beyond epoch 34, the improvements in both training and validation metrics became marginal, indicating saturation in learning. The final epoch (epoch 50) saw a training F1-score of 0.9004, but the validation F1-score plateaued at 0.3704, with the validation loss peaking at 2.7500 by epoch 44, pointing to signs of overfitting. To enhance the model's generalization, further optimization is needed. Implementing regularization techniques, such as dropout layers and L2 weight decay, can help reduce overfitting. Additionally, data augmentation strategies, such as generating synthetic text samples, could expand the dataset and improve model robustness. Leveraging transfer learning by utilizing multilingual pre-trained models is another promising approach to improving feature extraction and enhancing model performance (Ruder et al., 2019). These refinements aim to improve classification accuracy and ensure better generalization for unseen Tulu text, addressing the limitations observed during training.

Test Scores. We achieved test F1-scores of 0.4299 for Tamil and 0.1546 for Tulu. These results demonstrate the model's varying effectiveness across different languages.

5 Discussion

5.1 Performance Disparity Between Tamil and Tulu

The large gap in F1-scores between Tamil (0.4299) and Tulu (0.1546) stems from differences in resource availability. Tamil benefits from a well-pretrained model (`l3cube-pune/indic-sentence-similarity-sbert`) and a larger corpus, allowing better semantic representation. Tulu, in contrast, suffers from data scarcity and lacks a dedicated pre-trained model, leading to poor generalization and lower classification performance.

5.2 Overfitting Issues

The significant gap between training and validation performance, particularly in Tulu, indicates overfitting. Due to limited training data, the model memorizes patterns rather than learning generalizable features. The high model complexity relative to the dataset size, along with weak regularization, exacerbates this issue. Future improvements should include data augmentation, stronger regularization, and transfer learning to enhance generalization.

5.3 Linguistic Challenges

Tulu faces several linguistic challenges, including its low digital presence, morphological complexity, and lack of a standardized script. The frequent use of code-mixing with Kannada further complicates text classification. Unlike Tamil, which has well-structured textual data, Tulu's inconsistencies make NLP tasks more difficult. Addressing these challenges requires better datasets, improved tokenization, and language-specific embeddings.

5.4 Future Directions

Future work should focus on transfer learning with related languages, data augmentation for enhanced training diversity, and improved regularization to reduce overfitting. Most importantly, developing Tulu-specific models will be crucial for advancing NLP research in low-resource languages, ensuring greater inclusivity in AI applications.

6 Conclusion

This research proposed a hybrid pipeline for text classification in Tamil and Tulu, combining SBERT embeddings with custom MLP architectures. Despite limited resources, our methodology achieved promising results, with validation F1-scores of 0.4218 for Tamil and 0.3935 for Tulu, demonstrating the framework's ability to capture semantic patterns. While overfitting and metric saturation posed challenges, the model's early-stage improvements and balanced performance underscore its potential. Future work, including data augmentation and transfer learning, can further enhance accuracy and generalization. This study lays a solid foundation for advancing text classification in multilingual and low-resource settings, providing valuable insights for developing robust models in similar linguistic environments.

Limitations

One limitation of this work is the reliance on small, potentially non-representative datasets, which can hinder the model's ability to generalize effectively across diverse real-world scenarios. While the model performed well in the initial stages, issues like overfitting, particularly in the Tulu language model, highlight the challenge of training deep learning models on limited data. Moreover, the absence of high-quality annotated resources for these languages remains a significant barrier to achieving optimal performance.

Broader Impact Statement

Despite these challenges, the broader impact of this research is substantial. By developing a hybrid pipeline for text classification in low-resource languages like Tamil and Tulu, the work lays a foundation for more inclusive AI applications. Enhancing language accessibility through AI can bridge gaps in sentiment analysis, content moderation, and social media monitoring, especially for regional languages. This research demonstrates the potential to create more equitable and culturally aware AI systems that can serve underrepresented linguistic communities worldwide.

Acknowledgement

We express our sincere gratitude to Computational Intelligence and Operations Laboratory (CIOL) for their invaluable guidance, unwavering support, and continuous assistance throughout this journey. We are deeply appreciative of their efforts in organizing the CIOL Winter ML Bootcamp (Wasi et al., 2024), which provided an enriching learning environment and a strong foundation for collaborative research. The research mentoring and structured support offered by CIOL played a pivotal role in shaping this work, fostering innovation, and empowering participants to contribute meaningfully to the field of computational linguistics.

References

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert. *arXiv preprint arXiv:2304.11434*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Mehrdad Farahani. 2021. Zabanshenas is a solution for identifying the most likely language of a piece of written text.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, SUBALALITHA CN, Lavanya S K, Thenmozhi D, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023a. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Asha Hegde, G Kavya, Sharal Coelho, Pooja Lamani, and Hosahalli Lakshmaiah Shashirekha. 2023b. Munlp@ dravidianlangtech2023: Learning approaches for sentiment analysis in code-mixed tamil and tulu text. In *Proceedings of the Third Workshop*

on Speech and Language Technologies for Dravidian Languages, pages 275–281.

Soon-Jae Hwang and Chang-Sung Jeong. 2023. Integrating pre-trained language model into neural machine translation. In *2023 2nd International Conference on Frontiers of Communications, Information System and Data Science (CISDS)*, pages 59–66. IEEE.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

David MW Powers. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18.

Lavanya S. K., Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 2024. Overview of Second Shared Task on Sentiment Analysis in Code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Azmine Toushik Wasi, MD Shakiqul Islam, Sheikh Ayatur Rahman, and Md Manjurul Ahsan. 2024. Ciol presnts winter ml bootcamp. 6 December, 2024 to 6 February, 2025.

Xue Ying. 2019. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing.

KEC_AI_DATA_DRIFTERS@DravidianLangTech 2025: Fake News Detection in Dravidian Languages

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2], **Vishali K S**[1], **Priyanka B**[1], **Naveen Kumar K**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{vishaliks.22aid, priyankab.22aid, naveenkumark.22aid}@kongu.edu

Abstract

Detecting fake news in Malayalam possess significant challenges due to linguistic diversity, code-mixing, and the limited availability of structured datasets.We participated in the Fake News Detection in Dravidian Languages shared task, classifying news and social media posts into binary and multi-class categories. Our experiments used traditional ML models: Support Vector Machine (SVM), Random Forest, Logistic Regression, Naive Bayes and transfer learning models: Multilingual Bert (mBERT) and XLNet. In binary classification, SVM achieved the highest macro-F1 score of 0.97, while in multi-class classification, it also outperformed other models with a macro-F1 score of 0.98. Random Forest ranked second in both tasks. Despite their advanced capabilities, mBERT and XLNet exhibited lower precision due to data limitations. Our approach enhances fake news detection and NLP solutions for low-resource languages.

1 Introduction

Fake news detection is a critical challenge in the digital age, where misinformation spreads rapidly online, influencing public opinion and policy-making. This issue is even more pronounced in regional languages like Malayalam due to linguistic complexities, cultural nuances, and diverse online content. Developing effective detection methods is essential to ensure informed decision-making.

Detecting fake news in low-resource languages like Malayalam is challenging due to limited annotated data, linguistic diversity, and writing style variations like code-mixing and Romanization. Additionally, Malayalam's complex morphology and informal online discourse make classification difficult. Existing approaches primarily use machine learning and transfer learning techniques to improve classification accuracy.

This research develops a fake news detection system for Malayalam by evaluating traditional machine learning models and transformer-based approaches independently. We applied SVM, Random Forest, Multinomial Naive Bayes, and Logistic Regression for text classification. Additionally, transfer learning models such as mBERT and XLNet were tested to analyze multilingual text. Each model's effectiveness was assessed separately to determine the most suitable approach for fake news classification in Malayalam.

2 Literature Review

Subramanian et al. (2024a) highlighted how social networks spread misinformation, affecting public understanding and trust. The FakeDetect-Malayalam shared task addresses this with two subtasks: Task 1 classifies social media posts as genuine or fake, while Task 2 categorizes fake news into five labels, including False and Mostly True.

Subramanian et al. (2024b) highlighted the rapid spread of fake news in Malayalam online. In Task 1, they secured ninth place using RNNs to classify news as Original or Fake, leveraging their ability to capture sequential patterns. Their study aims to enhance accuracy and improve fake news detection.

Qu et al. (2024) proposed QMFND, a quantum fusion model using Quantum Convolutional Neural Networks (QCNN) to merge text and image data. Tested on Gossip and Politifact datasets, it shows robustness against quantum noise, enhances expressibility, and performs as well as or better than classical models.

K et al. (2024) emphasized the need for fake news detection in Malayalam, a low-resource language. They introduced a curated dataset categorizing news by inaccuracy levels. Baseline models, including multilingual BERT and ML classifiers, showed potential, with Logistic Regression on LaBSE achieving an F1 score of 0.3393. Addressing data imbalance is key to improving accuracy.

Shanmugavadivel et al. (2024) emphasized the rapid spread of false information on social media and the need for fake news detection. The study validated YouTube comments using ML models like Naive Bayes, SVM, Random Forest, and Decision Tree. Presented at DravidianLangTech@EACL 2024, it applies ML and NLP techniques to combat misinformation.

Babu et al. (2023) proposed a machine learning approach for vectorizing and tokenizing news headlines. Experimental results demonstrate that this method surpasses existing fake news detection techniques, demonstrating its effectiveness in various topics and languages.

S et al. (2023) provided detailed instructions for preparing a manuscript for the RANLP 2023 proceedings on this page. These guidelines apply to both initial submissions and final versions, including an example of the required format. The authors must follow all provided instructions to ensure compliance.

Hu et al. (2022) explored deep learning approaches for fake news detection, including supervised, weakly supervised, and unsupervised methods. The study evaluates models using news content, social context, and external data while reviewing FND datasets, identifying limitations, and proposing future research directions.

Baarir and Djeffal (2021) proposed a machine learning system for fake news detection using TF-IDF with bag of words, n-grams for feature extraction, and SVM for classification. The system was trained on a curated dataset of true and fake news, demonstrating its effectiveness in identifying misinformation. However, detecting fake news remains challenging due to limited datasets and analytical approaches.

Ahmad et al. (2020) developed a machine learning ensemble model to classify news articles based on linguistic characteristics. Evaluated on four real-world datasets, the model demonstrated superior performance compared to individual classifiers in detecting disinformation.

3 Problem and System description

The Fake News Detection from Malayalam News task aims to identify fake news in Malayalam social media posts and articles. It classifies the text into two categories: Fake or original, with the dataset containing labeled posts and articles. The task also includes categorizing fake news into five labels: False, Half True, Mostly False, Partly False, and Mostly True. Researchers will use machine learning models, embeddings, and transfer learning to distinguish between accurate and misleading information. This task contributes to creating a robust fake news detection system for Malayalam content, promoting reliable communication, and reducing misinformation. Subramanian et al. (2025) Out of 128 teams, our system secured the 13th rank in task 1 and the 5th rank in task 2.

4 Dataset description

The shared dataset consists of Malayalam news articles categorized into two tasks. In Task 1 (Binary Classification), the training dataset consists of 1,659 Original and 1,598 Fake class labels, while the test dataset consists of 1,019 rows. In Task 2 (Multiclass Classification), the training dataset consists of 1,384 False, 2 True, 162 Half True, 295 Mostly False, and 57 Partly False labels, with a test dataset containing 200 rows. In addition, a validation dataset is provided to evaluate the performance of the model prior to testing.

Dataset	No. of Comments
Train	3257
Dev	815
Test	1019

Table 1: Task 1 Dataset Description

Dataset	No. of Comments
Train	1900
Test	200

Table 2: Task 2 Dataset Description

5 Methodology

5.1 Data pre-processing

To enhance text quality for fake news classification, we applied a structured pre-processing pipeline. The text was lowercased, and URLs, HTML tags, special characters, and numbers were removed. Tokenization filtered out non-informative words using custom Malayalam stopwords. Stemming (PorterStemmer) and lemmatization (WordNetLemmatizer) normalized words, while short and duplicate words were removed to reduce redundancy. This process improves data quality for machine learning analysis. Figure 1 illustrates the workflow from data pre-processing to classification.

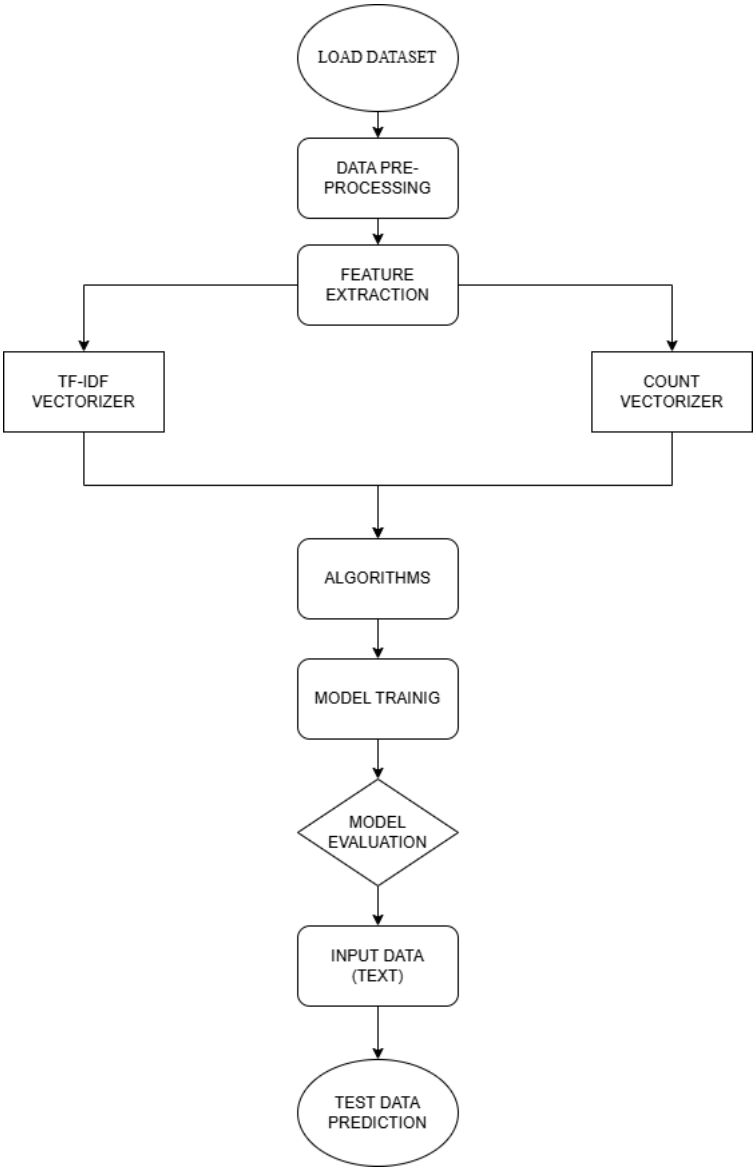


Figure 1: Proposed System Workflow

5.2 Encoding module

For our dataset, we utilized TF-IDF Vectorizer and Count Vectorizer from sklearn.feature-extraction for feature extraction. TF-IDF assigns weights to words based on their importance, reducing the influence of commonly used terms. Count Vectorizer, on the other hand, generates a matrix representing word frequencies, highlighting frequently occurring words in the text. By combining both techniques, we ensure a balanced representation of important terms and common patterns. This approach enhances the model's ability to capture textual features effectively, leading to improved classification performance. Ultimately, these methods contribute to better accuracy in fake news detection.

5.3 Model description

To classify Malayalam news as original or fake, we used SVM, Random Forest, Multinomial Naive Bayes, and Logistic Regression for text classification. Naive Bayes assigns probabilities based on Bayes' theorem, Random Forest builds multiple decision trees, SVM finds an optimal hyperplane, and Logistic Regression predicts binary outcomes. Additionally, XLNet and mBERT were applied for transfer learning, with XLNet leveraging bidirectional context and mBERT supporting multilingual analysis. Each model was tested independently, and results showed that traditional ML models outperformed transformer-based approaches for fake news detection in Malayalam.

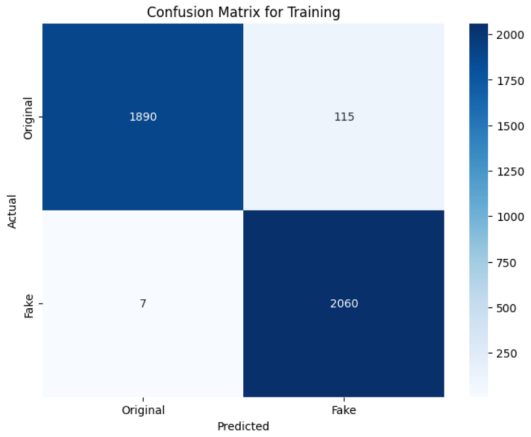


Figure 2: Confusion Matrix for Task 1

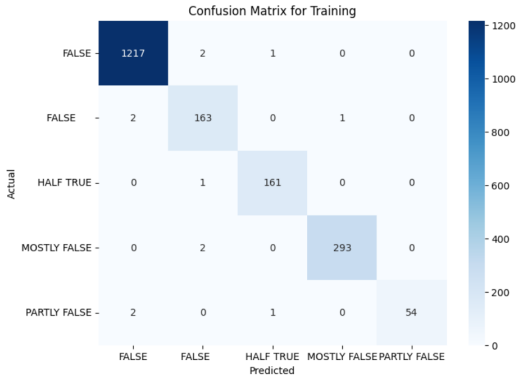


Figure 3: Confusion Matrix for Task 2

6 Experimental Analysis

In this experiment, we used the Malayalam dataset for the classification of fake news, applying four machine learning models and two transfer learning models for two tasks. In task 1 (binary classification), SVM achieved the highest accuracy (97%), followed by Random Forest (96%), Logistic Regression (94%), Multinomial Naive Bayes (90%), XLNet (80%), and mBERT (83%). SVM was the best performer, excelling in both accuracy and macro F1 score. In task 2 (multiclass classification), SVM again led with 98% accuracy, followed by Random Forest (95%), Logistic Regression (92%), and Multinomial Naïve Bayes (91%). Traditional models, especially SVM and Random Forest, outperformed deep learning models like XLNet and mBERT, proving to be the most reliable for fake news detection in Malayalam. Github Repository: Fake News Detection

7 Limitations

Our approach relies heavily on labeled datasets, which are limited to Malayalam and other Dravidian languages. The imbalance in multi-class classification affected model performance, especially for underrepresented labels. Although traditional ML models performed well, transfer learning models struggled due to data scarcity and domain-specific challenges. Furthermore, variations in code-mixed and informal text reduced the accuracy of the classification. Enhancing dataset quality, incorporating advanced preprocessing techniques, and optimizing deep learning models are crucial to improve fake news detection in Malayalam.

Model	Macro F1-Score
Support Vector Classifier	0.97
Random Forest	0.96
Logistic Regression	0.94
Naive Bayes	0.90
mBert	0.83
XLNet	0.80

Table 3: Macro F1-Score Metrics for Task 1

Model	Macro F1-Score
Support Vector Classifier	0.98
Random Forest	0.95
Logistic Regression	0.92
Naive Bayes	0.91

Table 4: Macro F1-Score Metrics for Task 2

8 Conclusion

We applied multiple machine learning and transfer learning models for fake news detection in Malayalam. SVM and Random Forest performed exceptionally well, achieving high Macro-F1 scores in both tasks, with SVM scoring 0.97 in binary and 0.98 in multi-class classification. Although mBERT and XLNet had lower accuracy due to data constraints, they demonstrated the potential of context-aware models for low-resource languages. Each model was applied independently to assess its effectiveness, and the results highlight the superiority of traditional ML models over transformer-based approaches in this context. Future work should focus on improving the quality of the data set, improving feature extraction techniques and optimizing deep learning models to further advance fake news detection in Malayalam and other Dravidian languages.

References

Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. 2020. Fake news detection using machine learning ensemble methods. *Complexity*, 2020:1–11.

Nihel Fatima Baarir and Abdelhamid Djeffal. 2021. Fake news detection using machine learning. In *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, pages 125–130.

Tina Babu, Rekha R Nair, Adithya Challa, Rahul Srikanth, Sri Sai Aravindan, and Suhas S. 2023. Fake news detection using machine learning algorithms. In *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, volume 1, pages 1–7.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3:133–155.

Devika K, Hariprasath .s.b, Haripriya B, Vigneshwar E, Premjith B, and Bharathi Raja Chakravarthi. 2024. From dataset to detection: A comprehensive approach to combating Malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23, St. Julian's, Malta. Association for Computational Linguistics.

Zhiguo Qu, Yunyi Meng, Ghulam Muhammad, and Prayag Tiwari. 2024. Qmfnd: A quantum multimodal fusion-based fake news detection model for social media. *Information Fusion*, 104:102172.

Malliga S, Bharathi Raja Chakravarthi, Kogilavani S V, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, and Muskaan Singh. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 59–63, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kogilavani Shanmugavadivel, Malliga Subramanian, Sanjai R, Mohammed Sameer B, and Motheeswaran K. 2024. Beyond tech@DravidianLangTech2024 : Fake news detection in Dravidian languages using machine learning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 124–128, St. Julian's, Malta. Association for Computational Linguistics.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Premjith B, Vanaja K, Mithunja S, Devika K, Hariprasath S.b, Haripriya B, and Vigneshwar E. 2024a. Overview of the second shared task on fake news detection in Dravidian languages: DravidianLangTech@EACL 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78, St. Julian's, Malta. Association for Computational Linguistics.

Malliga Subramanian, Jayanthjr J R, Muthu Karuppan P, Keerthibala T, and Kogilavani Shanmugavadivel. 2024b. KEC_HAWKS@DravidianLangTech 2024 : Detecting Malayalam fake news using machine learning models. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 266–270, St. Julian's, Malta. Association for Computational Linguistics.

KECEmpower@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media

Malliga Subramanian[1], Kogilavani Shanmugavadivel[2], Indhuja V S[1], Kowshik P[1], Jayasurya S[1]
[1]Department of CSE, Kongu Engineering College, Perundurai, Erode.
[2]Department of AI, Kongu Engineering College, Perundurai, Erode.
{mallinishanth72, kogilavani.sv}@gmail.com
{indhujavs.23cse, kowshikp.23cse}@kongu.edu
jayasuryas.23cse@kongu.edu

Abstract

The detection of abusive text targeting women, especially in Dravidian languages like Tamil and Malayalam, presents a unique challenge due to linguistic complexities and code-mixing on social media. This paper evaluates machine learning models such as Support Vector Machines (SVM), Logistic Regression (LR), and Random Forest Classifiers (RFC) for identifying abusive content. Code-mixed datasets sourced from platforms like YouTube are used to train and test the models. Performance is evaluated using accuracy, precision, recall, and F1-score metrics. Our findings show that SVM outperforms the other classifiers in accuracy and recall. However, challenges persist in detecting implicit abuse and addressing informal, culturally nuanced language. Future work will explore transformer-based models like BERT for better context understanding, along with data augmentation techniques to enhance model performance. Additionally, efforts will focus on expanding labeled datasets to improve abuse detection in these low-resource languages.

1 Introduction

The rise of social media has led to an increase in online abuse, particularly gender-based harassment targeting women. Detecting such abusive content in languages like Tamil and Malayalam presents unique challenges for Natural Language Processing (NLP). Despite growing concerns, there is limited research on abusive language detection in Tamil and Malayalam languages. This study explores the effectiveness of machine learning models, including Support Vector Machines (SVM), Logistic Regression, and Random Forest, for identifying abusive content in Tamil and Malayalam social media posts.This paper contributes to advancing content moderation techniques for multilingual social media platforms.

2 Literature Survey

Recent studies on abusive language detection have predominantly concentrated on English, yielding promising results with advanced machine learning models. However, research on low-resource languages, particularly Dravidian languages such as Tamil and Malayalam, remains limited. These languages often feature code-mixing, informal expressions, and context-dependent nuances, posing unique challenges for accurate detection Priyadharshini et al., 2022. The survey provides an overview of models submitted for abusive text identification in DravidianLangTech@NAACL 2025. Additionally, the absence of large annotated datasets and the difficulty of detecting implicit and subtle abuse further complicate the task Chen et al., 2018. Future research should focus on developing robust, language-specific models and expanding annotated datasets to enhance detection accuracy.

2.1 Abusive Detection in English and Major Language

Early abusive content detection relied on blacklists and regular expressions but struggled with subtle expressions, sarcasm, and context-dependent abuse Jiangbin et al., 2021. Machine learning models improved detection using features like n-grams and sentiment analysis Akhter et al., 2022, yet they struggled with nuanced language. Transformer-based models like BERT, RoBERTa, and ALBERT enhanced accuracy by capturing context through self-attention mechanisms.

2.2 Abusive Detection in Dravidian Languages

Abusive language detection in Tamil and Malayalam, especially gender-targeted abuse on social media, remains under-explored despite its importance. These Dravidian languages pose challenges due to complex syntax, rich morphology, and

limited annotated datasets. Traditional models like SVM and Random Forest struggle with implicit, code-mixed, and culturally nuanced abuse Eshan and Hasan, 2017. While deep learning and transformer models show promise, they still face difficulties in capturing subtle abuse. There is a pressing need for culturally aware models and expanded datasets to improve detection in these languages.

2.3 Deep Learning and Transformers

Recent advancements in deep learning, especially with transformer-based models like BERT, RoBERTa, and ALBERT, have significantly enhanced the accuracy of abusive language detection in social media Rajiakodi et al., 2025. These models are particularly powerful in handling the complexity of informal, code-mixed, and context-dependent expressions in Tamil and Malayalam. Unlike traditional methods, which rely on handcrafted features, transformers utilize self-attention mechanisms to understand the relationship between words in a sentence. This enables them to capture subtle forms of abuse such as implicit and gender-targeted language. However, the lack of large annotated datasets in these languages remains a challenge. Fine-tuning transformer models on Tamil and Malayalam data is crucial to improving performance Priyadharshini et al., 2023.

3 Materials and Methods

3.1 Taskset Description

This study focuses on detecting abusive language in Tamil-English and Malayalam text, particularly targeting women, sourced from social media platforms such as YouTube. The dataset was collected by scraping YouTube comments using specific queries related to controversial and sensitive topics, ensuring a diverse representation of abusive and non-abusive content. Figures 1 and 2 provide sample datasets for Tamil and Malayalam, respectively. The dataset consists of 5,723 texts in the training set, 1,229 texts in the development set, and 1,227 texts in the test set, each labeled as "Abusive" or "Non-Abusive." These texts exhibit significant linguistic challenges, including code-mixing, informal expressions, and culturally nuanced content, which are common in Dravidian languages. The data was processed and annotated to facilitate robust classification, leveraging machine learning techniques. By employing models like SVM, Random Forest, and Logistic Regression, this study evaluates classification performance across diverse linguistic patterns.

Text	Class
You tube ல இப்படி எல்லாம் முன்னேற்றம் நம் சேனலையும வந்து முன்னேற்றுங்கள்!!	Non-Abusive
ஆமா பா கட்டி வச்சி தோல உரிக்கணும	Abusive
சிறியவர்கள், இதயநோயாளிகள் , கர்ப்பிணி பெண்கள் யாரும் இதை பார்க்க வேண்டாம்	Non-Abusive
உன் பல்ல பாத்தாலே பயமா இருக்கு. ஆளும் பல்லும் உருவமும் பேச்சும் பாரு. இதுல 2 kg டூ வேற	Abusive
இவயா இதற்கெல்லாம் அர்த்தங்கள் என்ன கேக்க வேண்டும் உன்னை காலம் கை கூடினால்.....	Non-Abusive

Figure 1: Sample training texts from Tamil dataset

Text	Class
തൊലിക്കട്ടി ഉണ്ടോ എന്നാൽ എന്ത് പിഴവ് വേണമെങ്കിലും ഉണ്ടാക്കാൻ പറ്റും	Abusive
ആ കുലയുടെ രണ്ട് വിത്ത് കിട്ടുമോ ?	Abusive
അടി തെറ്റിയാൽ .... ജിമിക്കിയും കമ്മലും വിഴും	Non-Abusive
"എന്ത് നാണക്കേട്.ആല് കിളിത്തലും അത് തണലായി കരുതുന്ന ടീംസ് ആണ്."	Abusive
"നോട്ടപ്പിശക് പോലും കേൾക്കുമ്പോൾ കലി വരുന്നു. ഒടുക്കത്തെ നോട്ടപ്പിശക്"	Non-Abusive

Figure 2: Sample training texts from Malayalam dataset

3.2 Preprocessing and Feature Extraction

Preprocessing plays a crucial role in transforming raw text data into a format suitable for machine learning. The dataset was first cleaned by handling missing values and removing incomplete rows. The labels for "Abusive" and "Non-Abusive" content were mapped to binary values (1 and 0) for binary classification. For feature extraction, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique was applied, converting text into numerical features while capturing the significance of words. Stop words were eliminated to remove common, non-informative words that would contribute little to classification. This process reduced noise and enhanced the model's ability to focus on relevant terms Sai and Sharma, 2021. Additionally, the text data was standardized and formatted to ensure uniformity. While the current approach focuses on TF-IDF-based vectorization, future improvements could integrate transliteration normalization through mapping dictionaries and custom tokenization strategies to better handle transliterated words and mixed scripts. These preprocessing steps ensured that the models received high-quality input, enabling efficient training and accurate predictions for abusive language detection in social media.

3.3 Models and Methodology

This study explores the use of machine learning models to detect abusive language in Tamil and Malayalam text. We employed three models: Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest Classifier (RFC). SVM, with its ability to find optimal decision

boundaries, was used for precise classification. Logistic Regression, a linear model, was utilized for its interpretability and effectiveness in binary classification tasks. Random Forest, an ensemble method, leveraged multiple decision trees to enhance predictive performance Mahmud et al., 2024. Text data was preprocessed using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique, which converts text into numerical features. The TF-IDF vectorizer was set to a maximum of 5000 features, with stop words removal enabled. The selected hyperparameters included a linear kernel for SVM, 100 estimators for Random Forest, and a random state of 42. The models were evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score.

4 Results and Discussion

The study on abusive content detection in Tamil and Malayalam revealed that while traditional machine learning models like Logistic Regression and Random Forest performed effectively, Support Vector Machine (SVM) demonstrated superior performance, making it the most suitable model for this task.

4.1 Performance Metrics

Model performance was evaluated using Accuracy, Precision, Recall, and F1-Score. Accuracy measures overall correctness, while Precision indicates correctly predicted abusive texts. Recall reflects the proportion of actual abusive texts identified, and F1-Score balances Precision and Recall, crucial for imbalanced datasets. These metrics help assess real-world effectiveness and optimize abuse detection systems. Given linguistic complexity of Tamil and Malayalam, they provide insights into model adaptability across diverse text patterns. Table 1 illustrates classification performance for Tamil, while Table 2 presents results for Malayalam, ensuring robust and reliable abuse detection models.

Classifiers	Class Labels	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	Non-Abusive	68	68	70	69
Logistic Regression	Abusive	68	70	68	69
Random Forest	Non-Abusive	68	67	69	68
Random Forest	Abusive	68	69	67	68
SVM	Non-Abusive	69	69	67	68
SVM	Abusive	69	69	71	70

Table 1: Performance of Classifiers for Abusive and Non-Abusive Text Detection in Tamil

Classifiers	Class Labels	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	Non-Abusive	64	63	67	65
Logistic Regression	Abusive	64	66	62	64
Random Forest	Non-Abusive	61	59	67	63
Random Forest	Abusive	61	64	56	60
SVM	Non-Abusive	64	63	67	65
SVM	Abusive	64	67	62	64

Table 2: Performance of Classifiers for Abusive and Non-Abusive Text Detection in Malayalam

4.2 Error Analysis

The SVM model performed well but struggled with indirect and implicit abusive language, often misclassifying it as non-abusive due to limited context understanding. It also faced challenges with Tamil-Malayalam code-mixed sentences, as TF-IDF failed to capture nuanced language patterns, leading to false negatives. Figure 5 provides an example of such misclassification. These limitations highlight the need for more context-aware models like BERT. Future work will focus on expanding the dataset, improving tokenization, and addressing ethical considerations for unbiased predictions.

id	Text	True Class	Predicted Class
14	கரும்பு இதுங்க மூஞ்சில முழிச்சா சோறு கெடைக்குமா?	Abusive	Non-Abusive
2	இப்படியே பேசிக்கிட்டே இருந்தா எப்டி.... யாரு பெருசுனு அடிக்கிகாட்டு ...	Non-Abusive	Abusive
137	ഞങ്ങളെപ്പറ്റി നല്ല വെറുത്ത ആണെങ്കിൽ ആവും ആ സുഖമെന കൂടെ ഇതുപോലെ ലെന്ത് തേച്ചു ഒട്ടിക്	Abusive	Non-Abusive
163	ഞാൻ വാലുപ്പ കൊടുക്കുന്നത്.... കേട്ട് കേട്ട് മടുത്തു	Abusive	Non-Abusive

Figure 5: Example of a misclassified Tamil-Malayalam code-mixed text by the SVM model.

5 Limitations

Abusive language detection in Tamil and Malayalam using traditional machine learning models such as Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM) encounters several hurdles. The scarcity of labeled datasets in these languages limits the models' effectiveness. Additionally, the intricate grammatical structures and diverse vocabulary present in Tamil and Malayalam complicate the accurate identification of offensive content. Existing models often rely on simple feature extraction techniques, which struggle to grasp the subtleties of code-mixing and contextual variations in online discourse. Lastly, recognizing gender-based abuse requires a deeper understanding of cultural context, which current models fail to fully address.

6 Conclusion

In this study, we explored the detection of abusive content targeting women in Tamil and

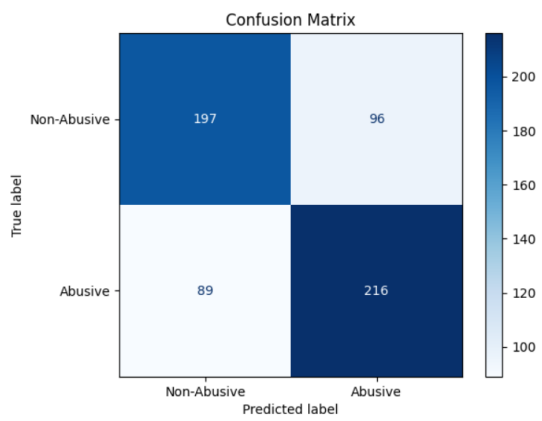


Figure 3: Confusion matrix for the Tamil dataset

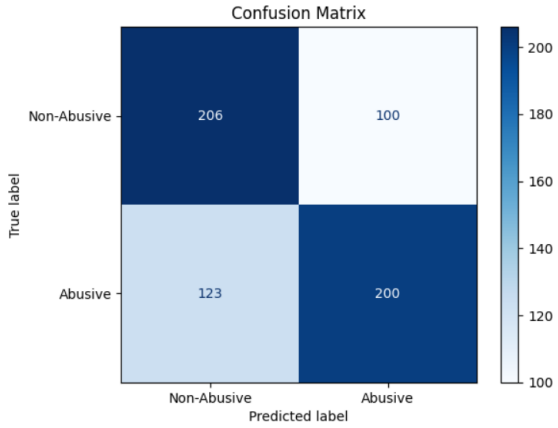


Figure 4: Confusion matrix for the Malayalam dataset

Malayalam social media posts using machine learning. Support Vector Machine (SVM) demonstrated superior performance in handling informal, context-dependent, and code-mixed language. The findings highlight challenges in detecting abuse in morphologically rich, low-resource languages and the need for scalable solutions. While results are promising, future research should focus on larger annotated datasets and hybrid models combining traditional methods with transformer-based architectures. The dataset and implementation code utilized in this study are publicly available at GitHub Repository to support reproducibility and further research.

References

M.P. Akhter, Z. Jiangbin, I.R. Naqvi, and et al. 2022. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimedia Systems*, 28:1925–1940.

H. Chen, S. McKeever, and S.J. Delany. 2018. A comparison of classical versus deep learning techniques for abusive content detection on social media sites. In *Social Informatics. SocInfo 2018*, volume 11185, pages 1–10. Springer, Cham.

S. C. Eshan and M. S. Hasan. 2017. An application of machine learning to detect abusive bengali text. In *Proceedings of the 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6, Dhaka, Bangladesh.

Zheng Jiangbin, Syed Irfan Naqvi, Mohammed Abdelmajeed, and Tehseen Zia. 2021. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimedia Systems*, 28.

T. Mahmud, T. Akter, M. K. Uddin, M. T. Aziz, M. S. Hossain, and K. Andersson. 2024. Machine learning techniques for identifying child abusive texts in online platforms. In *Proceedings of the 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6, Kamand, India.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the shared task on abusive tamil and malayalam text targeting women on social media: Dravidianlangtech@naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Siva Sai and Yashvardhan Sharma. 2021. Towards offensive language identification for dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 18–27, Kyiv. Association for Computational Linguistics.

KEC_AI_GRYFFINDOR@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages

Kogilavani Shanmugavadivel[1],**Malliga Subramanian**[2],
ShahidKhan S[1], **Shri Sashmitha S**[1], **Yashica S**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{shahidkhans.22,shrisashmithas.22,yashicas.22aid}@kongu.edu

Abstract

It is difficult to detect hate speech in code-mixed Dravidian languages because the data is multilingual and unstructured. We took part in the shared task to detect hate speech in text and audio data for Tamil, Malayalam, and Telugu in this research. We tested different machine learning and deep learning models such as Logistic Regression, Ridge Classifier, Random Forest, and CNN. For Tamil, Logistic Regression gave the best macro-F1 score of 0.97 for text, whereas Ridge Classifier was the best for audio with a score of 0.75. For Malayalam, Random Forest gave the best F1-score of 0.97 for text, and CNN was the best for audio (F1-score: 0.69). For Telugu, Ridge Classifier gave the best F1-score of 0.89 for text, whereas CNN was the best for audio (F1-score: 0.87).Our findings prove that a multimodal solution efficiently tackles the intricacy of hate speech detection in Dravidian languages. In this shared task,out of 145 teams we attained the 12th rank for Tamil and 7th rank for Malayalam and Telugu.

1 Introduction

The rise of social media has facilitated global communication but also led to the spread of hate speech. Detecting and preventing hate speech is crucial for fostering a safe and inclusive online space. This challenge intensifies in multilingual and code-mixed environments, such as Tamil, Malayalam, and Telugu, where users blend local scripts with borrowed words. The complexity of these languages, along with limited annotated datasets, makes hate speech detection a vital yet challenging research area.

Multimodal approaches combining text and audio offer deeper context for understanding online speech. While text-based models analyze linguistic cues, audio models capture tonal and prosodic features to detect aggression or hostility. This study employs machine learning and deep learning techniques, including Logistic Regression, Ridge Classifier, Random Forest, and CNN, to classify hate speech data from YouTube. By integrating both modalities, the methodology addresses limitations of conventional approaches.

Findings indicate that multilingual models can accurately detect hate speech across languages. Logistic Regression and Random Forest performed well in text classification, while CNNs effectively processed audio data. The results underscore the importance of combining linguistic and acoustic features to enhance detection accuracy. By expanding the multimodal dataset for Dravidian languages, this study contributes to building robust frameworks for combating hate speech in multilingual social media.

2 Literature Survey

Rawat et al. (2024) proposed a deep NLP model combining convolutional and recurrent layers for hate speech detection on social media, achieving a macro F1 score of 0.63 on the HASOC2019 dataset. The study also explored using unlabeled data and similar corpora to improve performance and reduce overfitting. Anbukkarasi and Varadhaganapathy (2023) introduced a synonym-based Bi-LSTM model to classify hate and non-hate texts in Tamil-English code-mixed tweets using a newly designed dataset of 10,000 annotated texts, addressing challenges of limited data and code-mixed language patterns.

As part of a collaborative effort, Tash et al. (2024) investigated Tamil hate speech detection related to migration and shelter, achieving an F1 score of 0.76 with a CNN model. Premjith et al. (2023) summarized a multimodal abusive language detection and sentiment evaluation effort in Tamil and Malayalam using video, audio, and text. The findings highlighted the challenges in creating ef-

fective models, with results based on the macro F1-score. Poornachandran et al. (2022) emphasized evaluating regional languages like Malayalam for hate speech detection, achieving an F1 score of 0.85 with deep learning techniques on a natural Malayalam dataset.

Priyadharshini et al. (2023) presented findings on abusive remark detection in Tamil and Telugu code-mixed social media text at RANLP 2023. The project developed models evaluated using the macro F1-score. Sai et al. (2024) explored hate speech detection in Telugu-English code-mixed text for DravidianLangTech@EACL-2024, achieving a macro F1 score of 0.65, ranking 14th in the competition.

Premjith et al. (2024a) analyzed submissions for Hate and Offensive Language Detection in Telugu Codemixed Text (HOLD-Telugu) at DravidianLangTech 2024, evaluating models using the macro F1-score. Another shared project led by Premjith et al. (2024b) focused on sentiment analysis, abusive language detection, and hate speech detection in Tamil and Malayalam using multimodal data. Despite 39 participants, only two submitted results, evaluated by the macro F1-score. Sreelakshmi et al. (2024) explored multilingual transformer-based embeddings for detecting hate speech in CodeMix Dravidian languages. Their study on Kannada-English, Malayalam-English, and Tamil-English datasets found MuRIL embeddings with an SVM classifier performed best. The research also addressed class imbalance with a cost-sensitive approach and introduced a new annotated Malayalam-English CodeMix dataset extending HASOC 2021.

3 Task Description

This study investigates multimodal hate speech detection in Tamil, Malayalam, and Telugu using YouTube-sourced text and audio data. Hate speech is categorized into Gender, Political, Religious, and Personal Defamation subclasses. Text preprocessing involved Count Vectorizer and TF-IDF, while audio preprocessing extracted prosodic features. Logistic Regression, Ridge Classifier, Random Forest, and CNN were applied, with performance evaluated using the macro-F1 score Lal G et al. (2025). Among 145 teams, our system ranked 12th for Tamil and 7th for Malayalam and Telugu, demonstrating the effectiveness of integrating text and audio models for detecting hate speech in Dravidian languages.

4 Dataset Description

4.1 Text Data Description

The text dataset for Malayalam, Tamil, and Telugu categorizes records as Hate or Non-Hate. Hate includes content labeled under Gender (G), Political (P), Religious (R), and Personal Defamation (C), while Non-Hate ('N') contains content without harmful language. The training set includes 883 Malayalam, 1,397 Tamil, and 1,953 Telugu records, with smaller test sets. Table 1 details the distribution of Hate and Non-Hate classes across languages, designed for training models in hate speech detection across multilingual contexts.

Language	Non-Hate(N)	Hate(C,G,P,R)
Malayalam	406	477
Tamil	287	491
Telugu	198	175

Table 1: Dataset Description of Text-Train

4.2 Audio Data Desciption

The audio dataset is structured similarly to the text dataset, with recordings labeled as Non-Hate or Hate. Hate includes content categorized under Gender (G), Political (P), Religious (R), and Personal Defamation (C). The training set has 883 Malayalam, 509 Tamil, and 551 Telugu recordings, with smaller test sets. Table 2 shows the distribution of Hate and Non-Hate categories across languages. This dataset helps train models for multilingual hate speech detection.

Language	Non-Hate(N)	Hate(C,G,P,R)
Malayalam	406	477
Tamil	287	222
Telugu	198	353

Table 2: Dataset Description of Audio-Train

5 Methodology

5.1 Data Preprocessing

Text and audio data underwent modality-specific preprocessing. Text processing included removing images, URLs, punctuation, tokenization, stopword removal, and stemming or lemmatization, followed by vectorization using Count Vectorizer and TF-IDF. Audio preprocessing involved noise reduction, normalization, and segmentation, with prosodic features like pitch and energy extracted to capture

speech tone. This approach ensured high-quality inputs for modeling.

5.2 Model Development

Logistic Regression, Ridge Classifier, Random Forest, and CNN were used for text and audio classification due to their effectiveness with high-dimensional data. These models captured linguistic and tonal features and were trained independently for Tamil, Malayalam, and Telugu to handle language-specific nuances. Class balancing, hyperparameter tuning, and cross-validation ensured robust performance.

5.3 Workflow Integration

The workflow integrates text and audio to enhance hate speech detection accuracy. Both modalities were processed separately and fed into their respective models. The outputs were analyzed to classify hate speech into categories like Gender, Political, Religious, and Personal Defamation. Figure 1 illustrates the workflow, covering preprocessing to classification. This modular design allows future experimentation with additional features or models, ensuring a comprehensive approach to multimodal hate speech detection in Dravidian languages.

6 Performance Evaluation

The performance of the fashions was evaluated based totally at the Macro-F1 score, that is a broadly used metric for category responsibilities, especially in imbalanced datasets. The fashions were educated on each text and audio records for the Tamil, Malayalam, and Telugu languages, and the respective performances are mentioned underneath.

6.1 Tamil

For text classification, Logistic Regression achieved the highest Macro-F1 score of 0.97, demonstrating strong accuracy in identifying hate speech in Tamil. Count Vectorizer and TF-IDF effectively transformed text into numerical representations, enabling the model to distinguish between Hate and Non-Hate categories. This high score indicates the model's ability to capture linguistic patterns, particularly in Gender (G), Political (P), Religious (R), and Personal Defamation (C) hate speech. Figure 2 presents the confusion matrix for the best-performing model.

For audio, Ridge Classifier performed best with a Macro-F1 score of 0.75. While CNN captured temporal speech features well, its overall performance was lower than other classifiers. Ridge Classifier's success suggests that spectral features significantly enhance hate speech detection in Tamil speech.

6.2 Malayalam

For the text modality, Random Forest achieved the highest Macro-F1 score of 0.97, demonstrating excellent performance in classifying hate speech across subclasses like Gender, Political, Religious, and Personal Defamation. As an ensemble method, Random Forest effectively leveraged features extracted through various vectorization techniques, ensuring strong predictions. Figure 3 presents the confusion matrix for the best-performing Malayalam text model.

For audio, CNN attained a Macro-F1 score of 0.69. While lower than the text score, it still showed reasonable success in detecting tonal patterns in Malayalam hate speech. The model's limitations may stem from the complexity of processing prosodic features.

6.3 Telugu

For the text modality, Ridge Classifier achieved a Macro-F1 score of 0.89, demonstrating strong performance in detecting hate speech and distinguishing between Non-Hate and Hate subclasses. The effectiveness of TF-IDF and Count-based vectorized features contributed significantly to the model's success. Figure 4 presents the confusion matrix for the best-performing Telugu text model.

For audio, CNN attained a Macro-F1 score of 0.87, effectively capturing speech dynamics in Telugu. Its strong performance highlights CNN's ability to analyze speech patterns and detect aggression or hostility.

7 Limitations

Our approach relies heavily on labeled datasets, which are limited for Dravidian languages. The complexity of prosodic features and insufficient audio samples affected audio model performance. Class imbalance in the dataset may have impacted the model's ability to generalize effectively. Expanding datasets and refining models are essential for addressing these limitations.

8 Conclusion

We applied multimodal approaches to classify hate speech in Tamil, Malayalam, and Telugu using text

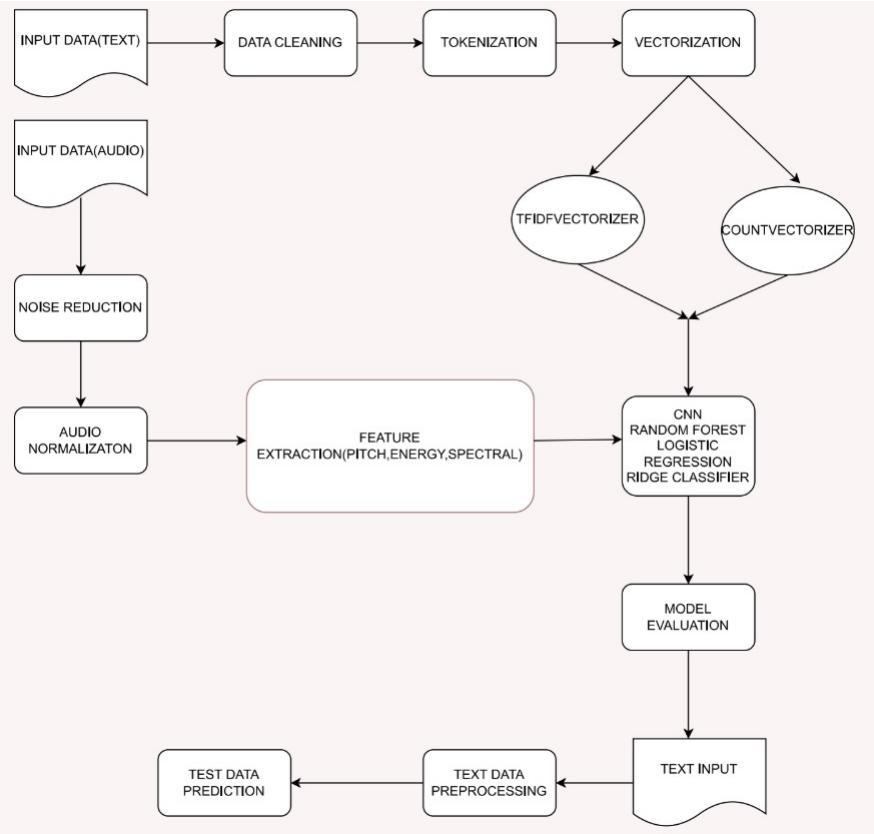


Figure 1: Proposed System Workflow

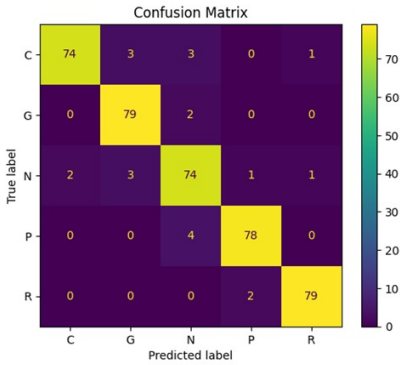


Figure 2: Confusion Matrix of Tamil-Text

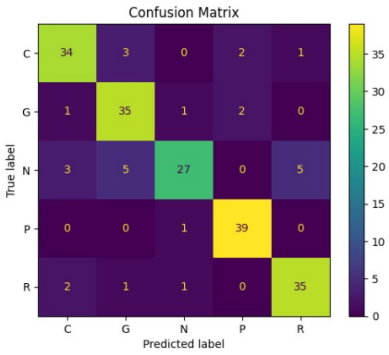


Figure 4: Confusion Matrix of Telugu-Text

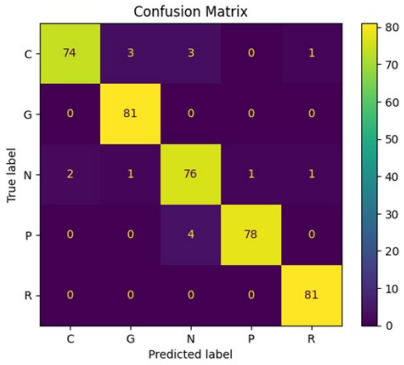


Figure 3: Confusion Matrix of Malayalam-Text

and audio data. Logistic Regression and Random Forest performed well for text, while CNN was most effective for audio, especially in Telugu. The results highlight the importance of combining linguistic and prosodic features for accurate detection. Overall, our approach shows promising results across languages. Further improvements in feature extraction and model optimization could enhance performance. The code for this shared task can be accessed at Github

References

S Anbukkarasi and S Varadhaganapathy. 2023. Deep learning-based hate speech detection in code-mixed tamil text. *IETE Journal of Research*, 69(11):7893–7898.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Prabaharan Poornachandran, VG Sujadevi, Gayathri Rajendran, Vinayak Ks, Vishnu Vijayan, Arjun Ram, et al. 2022. Malhate: Hate speech detection in malayalam regional language. In *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, volume 7, pages 110–115. IEEE.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

B Premjith, V Sowmya, Bharathi Raja Chakravarthi, Rajeswari Natarajan, K Nandhini, Abirami Murugappan, B Bharathi, M Kaushik, Prasanth Sn, et al. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Anchal Rawat, Santosh Kumar, and Surender Singh Samant. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(2):e1648.

Chava Sai, Rangoori Kumar, Sunil Saumya, and Shankar Biradar. 2024. Iiitdwd_svc@ dravidianlangtech-2024: Breaking language barriers; hate speech detection in telugu-english code-mixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 119–123.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

M Tash, Z Ahani, M Zamir, O Kolesnikova, and G Sidorov. 2024. Lidoma@ lt-edi 2024: Tamil hate speech detection in migration discourse. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 184–189.

KECLinguAIsts@DravidianLangTech 2025: Detecting AI-generated Product Reviews in Dravidian Languages

Malliga Subramanian[1], Rojitha R[1], Mithun Chakravarthy[1], Renusri R V[1], Kogilavani Shanmugavadivel[2]

[1]Department of CSE, Kongu Engineering College, Perundurai, Erode.
[2]Department of AI, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{rojithar.23cse, mithunchakravarthyy.23cse}@kongu.edu
renusrir.23cse@kongu.edu

Abstract

With the surge of AI-generated content in online spaces, ensuring the authenticity of product reviews has become a critical challenge. This paper addresses the task of detecting AI-generated product reviews in Dravidian languages, specifically Tamil and Malayalam, which present unique hurdles due to their complex morphology, rich syntactic structures, and code-mixed nature. We introduce a novel methodology combining machine learning classifiers with advanced multilingual transformer models to identify AI-generated reviews. Our approach not only accounts for the linguistic intricacies of these languages but also leverages domain-specific datasets to improve detection accuracy. For Tamil, we evaluate Logistic Regression, Random Forest, and XGBoost, while for Malayalam, we explore Logistic Regression, Multinomial Naive Bayes (MNB), and Support Vector Machines (SVM). Transformer-based models significantly outperform these traditional classifiers, demonstrating superior performance across multiple metrics.

1 Introduction

The rise of AI-generated content, particularly product reviews, has transformed online interactions but also raised concerns about authenticity. Detecting such reviews is especially challenging in low-resource languages like Tamil and Malayalam, which feature complex morphology, intricate syntax, and frequent code-mixing. Due to the scarcity of annotated datasets and linguistic diversity, developing robust detection models for these languages remains difficult. This study explores using traditional machine learning classifiers—Logistic Regression, Random Forest, and XGBoost for Tamil, and Logistic Regression, Multinomial Naive Bayes (MNB), and SVM for Malayalam—to detect AI-generated product reviews, demonstrating their effectiveness in addressing the unique challenges of low-resource languages.Chakravarthi et al., 2021

2 Literature Survey

Research on AI-generated product review detection has mainly focused on high-resource languages like English, with limited attention to low-resource Dravidian languages like Tamil and Malayalam. These languages pose challenges due to linguistic complexities such as code-mixing and context-dependent nuances. While some multilingual datasets exist, research specifically addressing AI-generated content in product reviews remains scarce.Baiju, 2023 The lack of large annotated datasets and the subtle nature of AI-generated reviews further complicate detection. This survey reviews models from DravidianLangTech@NAACL 2025 for AI-generated review detection. Advancements in transfer learning and multilingual embeddings could help address these challenges.Premjith et al., 2025

2.1 Detection of AI-Generated Content in English and Major Languages

Early methods for detecting AI-generated content relied on handcrafted features like text perplexity, word-level n-grams, and syntactic patterns. While effective for earlier AI systems, they struggled with modern generative models. Recent advancements, particularly transformer-based models like RoBERTa, GPT detectors, and BERT, enhance detection by analyzing contextual relationships and embedding patterns. These models identify subtle inconsistencies, such as repetitive phrasing and unnatural expressions, indicative of AI-generated text. However, their effectiveness is limited in Dravidian languages due to a lack of pre-trained models, emphasizing the need for fine-tuning on language-specific data.Muneer and Basheer, 2023

2.2 Detection of AI-Generated Reviews in Dravidian Languages

Detecting AI-generated product reviews in Tamil and Malayalam is an underexplored area, crucial due to the rising prevalence of such content online. Dravidian languages present challenges like complex syntax, rich morphology, and frequent code-mixing, complicating detection. The scarcity of annotated datasets for AI-generated content further limits model development. Current research mainly focuses on traditional text classification using Logistic Regression, Random Forest, and SVM, but these models struggle with subtle AI markers in code-mixed or informal text.Gautam and Bharathi, 2021 Transformer-based models like BERT and XLM-Roberta show promise but require extensive fine-tuning and dataset augmentation to address linguistic diversity.

2.3 Deep Learning and Transformers

Advancements in deep learning, particularly with transformer-based models like BERT, GPT, and XLM-R, have shown potential in addressing the challenges of detecting AI-generated reviews. These models use self-attention mechanisms to capture complex word relationships, enabling the detection of subtle AI-generated patterns. Unlike traditional methods relying on feature extraction, transformers analyze context holistically, making them effective for identifying generative AI outputs. However, their performance in Tamil and Malayalam is limited by the lack of large-scale annotated datasets.Sebastian, 2023 Fine-tuning them on Dravidian language-specific corpora is essential for better detection, especially in code-mixed and context-dependent reviews. Future improvements include using larger annotated datasets and transformer-based models like BERT or XLM-R to enhance contextual understanding and detection accuracy.

2.4 Practical Constraints of Transformer Models

Transformer models like BERT, mBERT, and XLM-R require significant computational power, making real-time deployment challenging. Their training time is much longer than traditional models, especially with multiple epochs. Large annotated datasets are essential, which are often scarce for low-resource languages like Tamil and Malayalam. Fine-tuning these models is complex, requiring careful hyperparameter tuning. Unlike traditional models, transformers lack interpretability, making their predictions harder to explain.

3 Materials and Methods

3.1 Taskset Description

This study focuses on detecting AI-generated product reviews in Tamil and Malayalam. The dataset consists of 1,608 training samples (808 Tamil, 800 Malayalam) and 300 test samples (100 Tamil, 200 Malayalam), each labeled as "AI" or "HUMAN", with a sample format shown in Figures 1 and 2. The task involves identifying whether a given review is AI-generated or human-written. Six machine learning models—Logistic Regression, Random Forest, XGBoost, Support Vector Machines (SVM), and Multinomial Naive Bayes (MNB)—are evaluated based on accuracy, precision, recall, and F1-score. The study aims to develop a robust framework for detecting AI-generated content in low-resource languages like Tamil and Malayalam.

ID	DATA	LABEL
TAM_HUAI_TR_386	இந்த கேபேஸ் வாஷ் சருமத்தை உலர்த்துகிறது.	AI
TAM_HUAI_TR_396	இந்த ஜ கிரீம் எரிச்சலை ஏற்படுத்துகிறது.	AI
TAM_HUAI_TR_401	இந்த நெயில் கபைல் விரலை கீறுகிறது.	AI
TAM_HUAI_TR_420	அதிகமா செலவு பண்ண வேண்டி இருக்கு	HUMAN
TAM_HUAI_TR_504	எனக்கு ஏற்ற அளவு செருப்பு கிடைக்கவில்லை	HUMAN

Figure 1: Sample training texts from Tamil dataset

ID	DATA	LABEL
MAL_HUAI_TR_012	കാജൽ റിമൂവറിന് ഈ പ്രോഡക്റ്റ് നല്ലതാണ്	HUMAN
MAL_HUAI_TR_021	നമുക്ക് ആവശ്യമായ സാധനങ്ങൾ അടിപൊളി	HUMAN
MAL_HUAI_TR_153	8 ആഴ്ച യൂസ് ചെയ്താൽ മതി. പെർമനൻറ് ആണ്	HUMAN
MAL_HUAI_TR_762	മീൻ കറിയിനോടൊപ്പം പപ്പടവും, പുളിശ്ശേരിയും പൊളിയാ!	AI
MAL_HUAI_TR_766	ഇവിടെ എവിടെ എങ്കിലും നാടൻ ഭക്ഷണവും ലഭിക്കും.	AI

Figure 2: Sample training texts from Malayalam dataset

3.2 Preprocessing and Feature Extraction

Preprocessing transformed the Tamil and Malayalam product review data into a machine learning-friendly format. The dataset, labeled as "AI" or "HUMAN", was cleaned by handling missing values, standardizing text, and converting it to lowercase. Stop words were removed, and TF-IDF vectorization was applied to extract numerical features. Tokenization and stemming were used to break text into words and reduce them to base forms. These steps ensured high-quality input for effective detection of AI-generated reviews.Sinthusha et al., 2025

3.3 Models and Methodology

This study uses machine learning models to detect AI-generated product reviews in Tamil and Malayalam text. For Tamil, we employed Logistic Regression (LR), Random Forest (RF), and XGBoost. LR was chosen for its interpretability, RF for its ensemble learning capabilities, and XGBoost for its ability to capture complex patterns. For Malayalam, we used Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Logistic Regression LR. SVM handles high-dimensional data, MNB is effective for word frequency distributions, and LR provides consistent performance.Priyadharshini et al., 2021 Text data was preprocessed using the TF-IDF vectorization technique, converting text into numerical features while handling missing values and code-mixing. The models were trained and evaluated based on accuracy, precision, recall, and F1-score to assess their effectiveness in detecting AI-generated reviews.

4 Results and Discussion

The study on detecting AI-generated product reviews in Tamil and Malayalam demonstrated that while various machine learning models, including Logistic Regression (LR), Random Forest (RF), and XGBoost, performed well, Logistic Regression outperformed the others in both languages. This makes Logistic Regression the most effective model for detecting AI-generated content in Tamil and Malayalam, showing its robustness in handling the intricacies of these low-resource languages. Figure 3 illustrates the Confusion Matrix for the high-performing model (LR) in Tamil and Malayalam.

4.1 Performance Metrics

The models were evaluated using Accuracy, Precision, Recall, and F1-Score. Accuracy measures the proportion of correctly classified reviews, while Precision indicates the percentage of correctly identified AI-generated reviews. Recall reflects the proportion of actual AI reviews detected, and F1-Score balances Precision and Recall, which is important for imbalanced datasets. These metrics are vital for assessing model performance. Table 1 shows the results on the training dataset, while Table 2 presents the test dataset performance, highlighting the variation in model effectiveness across datasets.

Classifiers	Class Labels	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (Tam)	AI	86	85	89	87
	HUMAN	86	88	84	86
Random Forest (Tam)	AI	91	88	94	91
	HUMAN	91	93	88	90
XGBoost (Tam)	AI	84	86	79	82
	HUMAN	84	83	88	85
Logistic Regression (Mal)	AI	74	77	70	73
	HUMAN	74	72	79	75
MNB (Mal)	AI	86	85	89	87
	HUMAN	86	88	84	86
SVM (Mal)	AI	79	78	80	79
	HUMAN	79	79	78	78

Table 1: Performance of Classifiers for AI and HUMAN Text Detection in Tamil and Malayalam(Training Dataset)

Classifiers	Class Labels	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (Tam)	AI	68	65	71	68
	HUMAN	68	71	65	68
Random Forest (Tam)	AI	68	67	67	67
	HUMAN	68	69	69	69
XGBoost (Tam)	AI	68	70	58	64
	HUMAN	68	67	77	71
Logistic Regression (Mal)	AI	67	68	64	66
	HUMAN	67	66	70	68
MNB (Mal)	AI	65	76	45	57
	HUMAN	65	61	86	71
SVM (Mal)	AI	65	67	62	64
	HUMAN	65	64	69	67

Table 2: Performance of Classifiers for AI and HUMAN Text Detection in Tamil and Malayalam(Test Dataset)

4.2 Error Analysis

Despite Logistic Regression achieving the highest accuracy, it misclassified several instances, predicting AI-generated reviews as HUMAN and vice versa. Figure 4 illustrates a few such examples, likely due to TF-IDF's inability to capture contextual nuances and challenges with code-mixed text.

ID	DATA	Predictions	Real_Predictions
TAM_HUAI_TE_086	பேக்கிங் சரியில்லை	AI	HUMAN
TAM_HUAI_TE_088	மிகவும் வாசனை உள்ள பொருள்	AI	HUMAN
MAL_HUAI_TE_104	ചുടുള്ള ഭക്ഷണം അലുമിനിയം ഫോയിലിൽ	HUMAN	AI
MAL_HUAI_TE_138	ഇക്കാലത്ത് ഒക്കെ ടാറ്റയുടെ ഡിസൈൻ കണ്ട്	HUMAN	AI

Figure 4: Example of a misclassified Tamil-Malayalam code-mixed text by LR model.

5 Limitations

Detecting AI-generated reviews in Tamil and Malayalam is challenging due to the lack of annotated datasets, limiting model performance. The complex syntax, rich morphology, and frequent code-mixing make feature extraction difficult. Traditional models relying on basic features like n-grams fail to capture subtle AI-generated patterns. Additionally, distinguishing human-like fluency in AI-generated content from human-authored text requires deeper linguistic understanding.Kumaresan and Pal, 2021 Cultural and contextual variations further complicate detection without robust language-specific datasets and tools.

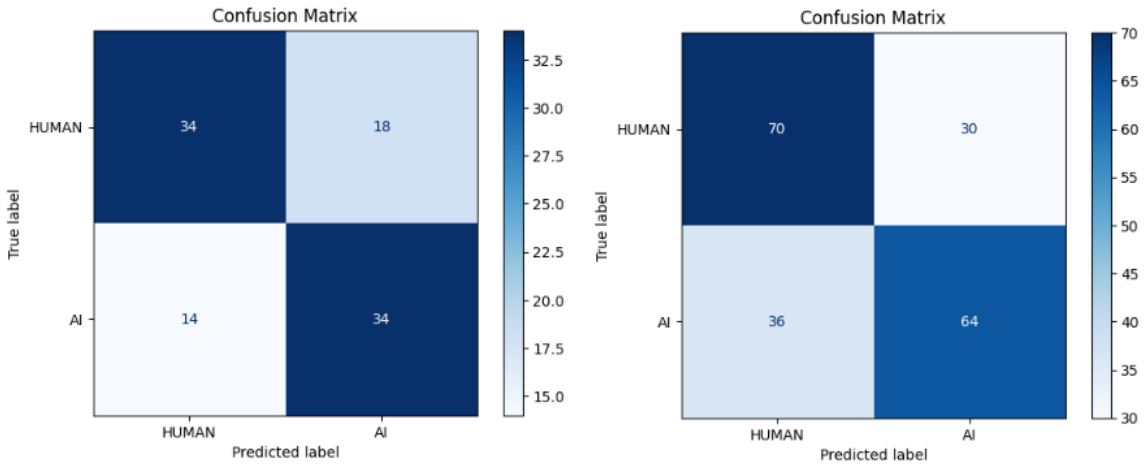


Figure 3: Confusion Matrix for high performing model(LR) in Tamil and Malayalam

6 Conclusion

This study explored detecting AI-generated and human-written reviews in Tamil and Malayalam using machine learning. Logistic Regression emerged as the most effective model, handling challenges in low-resource, morphologically rich languages.Nair et al., 2014 The results highlight the importance of tailored preprocessing and feature extraction techniques. Despite the promising outcomes, Future work should focus on creating larger annotated datasets and incorporating transformer-based models for better contextual understanding.Zhu and Dong, 2020 Hybrid approaches combining traditional and advanced models can further enhance detection accuracy in diverse linguistic contexts. The datasets and implementation code utilized in this research are publicly available at GitHub Repository to support reproducibility and further research.

References

K. B. Baiju. 2023. *Pattern primitive based malayalam handwritten character recognition studies for real-time applications*. Ph.D. thesis, Department of Computer Science, University of Calicut.

Bharathi Raja Chakravarthi et al. 2021. Dravidian-multimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*.

Abhishek Kumar Gautam and B. Bharathi. 2021. Rnn's vs transformers: Training language models on deficit datasets. In *FIRE (Working Notes)*, pages 737–743.

Kumar Kumaresan and Kingston Pal. 2021. Dravidian-multimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*.

V. K. Muneer and K. P. Mohamed Basheer. 2023. A collaborative destination recommender model in dravidian language by social media analysis. In *Proceedings of Data Analytics and Management: ICDAM 2022*, pages 541–551. Springer Nature Singapore.

Deepu S. Nair et al. 2014. Sentima-sentiment extraction for malayalam. In *2014 International conference on advances in computing, communications and informatics (ICACCI)*, pages 1719–1723. IEEE.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, Sajeetha Thavareesan, and Prasanna Kumar Kumaresan. 2025. Overview of the shared task on detecting ai generated product reviews in dravidian languages: Dravidianlangtech@naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruba Priyadharshini et al. 2021. Overview of the dravidiancodemix 2021 shared task on sentiment detection in tamil, malayalam, and kannada. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 4–6.

Mary Priya Sebastian. 2023. Malayalam natural language processing: challenges in building a phrase-based statistical machine translation system. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–51.

AV Ann Sinthusha, E. Y. A. Charles, and Ruvan Weerasinghe. 2025. Machine reading comprehension for the tamil language with translated squad. *IEEE Access*.

Yueying Zhu and Kunjie Dong. 2020. Yun111@ dravidian-codemix-fire2020: Sentiment analysis of dravidian code mixed text. In *FIRE (Working Notes)*, pages 628–634.

Dll5143@DravidianLangTech 2025: Majority Voting-Based Framework for Misogyny Meme Detection in Tamil and Malayalam

Sarbajeet Pattanaik
IIIT Allahabad
Prayagraj, 211015, India
mcl2023008@iiita.ac.in

Ashok Yadav
IIIT Allahabad
Prayagraj, 211015, India
rsi2021002@iiita.ac.in

Vrijendra Singh
IIIT Allahabad
Prayagraj, 211015, India
vrij@iiita.ac.in

Abstract

Misogyny memes pose a significant challenge on social networks, particularly in Dravidian-scripted languages, where subtle expressions can propagate harmful narratives against women. This paper presents our approach for the "Shared Task on Misogyny Meme Detection," organized as part of DravidianLangTech@NAACL 2025, focusing on misogyny meme detection in Tamil and Malayalam. To tackle this problem, we proposed a multi-model framework that integrates three distinct models: M1 (ResNet-50 + google/muril-large-cased), M2 (openai/clip-vit-base-patch32 + ai4bharat/indic-bert), and M3 (ResNet-50 + ai4bharat/indic-bert). The final classification is determined using a majority voting mechanism, ensuring robustness by leveraging the complementary strengths of these models. This approach enhances classification performance by reducing biases and improving generalization. Our model achieved an F1 score of 0.77 for Tamil, significantly improving misogyny detection in the language. For Malayalam, the framework achieved an F1 score of 0.84, demonstrating strong performance. Overall, our method ranked 5th in Tamil and 4th in Malayalam, highlighting its competitive effectiveness in misogyny meme detection.

1 Introduction

The rapid proliferation of social media has revolutionized communication, enabling individuals to share information, ideas, and opinions instantly (Yadav and Singh, 2025) (Yadav and Singh, 2024). However, this unprecedented connectivity has also led to a surge in harmful online content, including hate speech, trolling, and misogyny, which disproportionately targets women and other marginalized groups (Lin et al., 2024). Among the diverse forms of online expression, memes have emerged as a powerful yet controversial medium. Detecting misogyny in memes is a complex task that demands a nuanced understanding of multimodal data, as memes typically combine text and images to convey meaning (Suryawanshi et al., 2020). While misogyny meme detection in major languages like English has seen significant advancements, there is a pressing need to extend this effort to languages written in Dravidian languages like Tamil and Malayalam (Shaun et al., 2024). The introduction of datasets such as MDMD for Tamil and Malayalam memes (Ponnusamy et al., 2024) and benchmark data sets for the detection of misogynistic content (Gasparini et al., 2022) highlighted the importance of annotated resources. Key initiatives included the development of frameworks like MISTRA for misogynous meme classification (Jindal et al., 2024),and shared tasks such as (Chakravarthi et al., 2024). To address these challenges, the Shared Task on Misogyny Meme Detection is organized as part of DravidianLangTech@NAACL 2025. This task challenges participants to develop multimodal models for analyzing textual and visual elements of social media memes. We secured 5th in Tamil and 4th in Malayalam.

The paper is organized as follows: Section 2 covers task statistics and the dataset. Section 3 explains our methodology. Section 4 details the experimental setup and evaluation metrics, with Section 4.1 comparing model performance. Section 5 concludes our findings, and Section 6 discusses limitations and future improvements in misogyny meme detection for Tamil and Malayalam. Our implementation is available on GitHub. [1]

2 Task and Dataset Description

The Shared Task on Misogyny Meme Detection is organized as part of DravidianLangTech@NAACL 2025 (Chakravarthi et al., 2025). This task aims to

[1] https://github.com/Deeplearninglabiiita/DravidianLangTech-.git

advance multimodal machine learning techniques to identify misogyny in memes. Participants are tasked with designing innovative systems capable of comprehensively interpreting both the textual and visual elements of memes. By combining these modalities, the systems must accurately classify memes as either misogynistic or non-misogynistic. The unique aspect of this task lies in its focus on two Dravidian languages, Tamil and Malayalam, which adds a layer of complexity due to linguistic diversity, cultural nuances, and limited resources in these languages.

The datasets for this shared task were drawn from (Ponnusamy et al., 2024). Table 1 provides statistics on the dataset used for the shared task in Tamil and Malayalam..

Table 1: The statistics of the used dataset in Dravidian-LangTech@NAACL 2025

Category	Tamil		Malayalam	
	Misogyny	Non-misogyny	Misogyny	Non-misogyny
Train	285	851	259	381
Val	74	210	63	97
Test	89	267	78	122
Total	448	1328	400	600

3 Proposed Framework

We have proposed a framework that utilizes three distinct models: M1 (ResNet-50 + google/muril-large-cased), M2 (openai/clip-vit-base-patch32 + ai4bharat/indic-bert), and M3 (ResNet-50 + ai4bharat/indic-bert). Subsequently, we employed a majority voting mechanism for the final classification.

3.1 M1 (ResNet-50 + google/muril-large-cased)

In this study, we proposed a multimodal architecture to classify memes as **misogyny** or **non-misogyny** by leveraging both textual and visual information. The text modality is processed using the pre-trained MuRIL encoder. Given an input textual content T, it is first tokenized using MuRIL's tokenizer and then passed through the MuRIL encoder to extract contextualized embeddings:

$$\mathbf{T}_{\text{emb}} = \text{MuRIL}(T), \tag{1}$$

where $\mathbf{T}_{\text{emb}} \in \mathbb{R}^{n \times d}$, n is the number of tokens after padding or truncation, and d is the embedding dimension (768 for MuRIL). The embedding corresponding to the `[CLS]` token, which summarizes the entire text, is extracted and projected into a lower-dimensional feature space using a linear layer:

$$\mathbf{T}_{\text{feat}} = \text{ReLU}(\mathbf{W}_T \cdot \mathbf{T}_{\text{emb}}^{[CLS]} + \mathbf{b}_T), \tag{2}$$

where $\mathbf{T}_{\text{feat}} \in \mathbb{R}^{256}$ is the projected text feature, $\mathbf{W}_T \in \mathbb{R}^{256 \times 768}$ is a learnable projection matrix, and $\mathbf{b}_T \in \mathbb{R}^{256}$ is the bias term.

The visual modality is processed using ResNet-50, which is pre-trained on ImageNet. Each input image I is resized and normalized before passing through the ResNet-50 network. Features are extracted from the penultimate layer by removing the classification head:

$$\mathbf{I}_{\text{emb}} = \text{ResNet-50}(I), \tag{3}$$

where $\mathbf{I}_{\text{emb}} \in \mathbb{R}^{2048}$ represents the high-dimensional image embedding. To align the dimensionality of image and text features, a linear projection is applied to reduce the dimensionality:

$$\mathbf{I}_{\text{feat}} = \text{ReLU}(\mathbf{W}_I \cdot \mathbf{I}_{\text{emb}} + \mathbf{b}_I), \tag{4}$$

where $\mathbf{I}_{\text{feat}} \in \mathbb{R}^{256}$, $\mathbf{W}_I \in \mathbb{R}^{256 \times 2048}$, and $\mathbf{b}_I \in \mathbb{R}^{256}$ are learnable parameters. The features from both modalities are concatenated to form a joint multimodal representation:

$$\mathbf{F} = [\mathbf{T}_{\text{feat}}; \mathbf{I}_{\text{feat}}], \tag{5}$$

where $\mathbf{F} \in \mathbb{R}^{512}$ is the fused feature vector.

The fused feature vector $\mathbf{F}$ is passed through a classification head, which consists of a dropout layer followed by a dense layer for binary classification:

$$y = \sigma(\mathbf{W}_C \cdot \mathbf{F} + b_C), \tag{6}$$

where $\mathbf{W}_C \in \mathbb{R}^{1 \times 512}$, $b_C \in \mathbb{R}$, σ is the sigmoid activation, and $y \in [0, 1]$ represents the probability of the meme being misogynistic. To classify the meme, a threshold $\tau = 0.5$ is applied:

$$\hat{y} = \begin{cases} 1, & \text{if } y \geq \tau \text{ (Misogyny)}, \\ 0, & \text{otherwise (Non-Misogyny)}. \end{cases}$$

3.2 M2 (openai/clip-vit-base-patch32 + ai4bharat/indic-bert),

This sub-section presents a multimodal approach to classifying memes as misogynistic or non-misogynistic using text and image features. The model utilizes CLIP for image encoding and IndicBERT for textual feature extraction, followed by a classification layer.

Given an input text T, the text encoder (IndicBERT) extracts a feature representation:

$$\mathbf{h}_T = f_{\text{IndicBERT}}(T), \tag{7}$$

where $\mathbf{h}_T \in \mathbb{R}^{d_T}$ is the output embedding of the [CLS] token.

Given an image I, the CLIP model extracts a feature representation:

$$\mathbf{h}_I = f_{\text{CLIP}}(I), \tag{8}$$

where $\mathbf{h}_I \in \mathbb{R}^{d_I}$ represents the image embedding from the CLIP model.

The extracted text and image features are concatenated to form a fused representation:

$$\mathbf{h}_{\text{fused}} = [\mathbf{h}_T; \mathbf{h}_I], \tag{9}$$

where $[;]$ denotes concatenation, and $\mathbf{h}_{\text{fused}} \in \mathbb{R}^{d_T+d_I}$.

A fully connected layer maps the fused representation to a binary classification output:

$$\hat{y} = \sigma(\mathbf{W}\mathbf{h}_{\text{fused}} + \mathbf{b}), \tag{10}$$

where σ is the sigmoid activation function, $\mathbf{W}$ is the weight matrix, and $\mathbf{b}$ is the bias term. To classify the meme, a threshold $\tau = 0.5$ is applied and predicted label is given by:

$$y = \begin{cases} 1, & \hat{y} > 0.5, \\ 0, & \hat{y} \leq 0.5. \end{cases} \tag{11}$$

Binary cross-entropy loss is used to optimize the model:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right]. \tag{12}$$

3.3 M3 (ResNet-50 + ai4bharat/indic-bert)

In this model (M3), our approach leverages multimodal learning by combining textual and visual features using indic-bert and ResNet-50 respectively.

Each meme consists of an image and an associated text transcription. Given a dataset of N memes, each sample can be represented as:

$$M_i = (I_i, T_i, y_i) \tag{13}$$

where:

- $I_i \in \mathbb{R}^{H \times W \times C}$ is the meme image with height H, width W, and C color channels.
- T_i is the textual transcription.
- $y_i \in \{0, 1\}$ is the binary label (0: non-misogynistic, 1: misogynistic).

We use IndicBERT to encode text representations. Given a transcription T_i, the tokenized input is:

$$\mathbf{x}_i = \text{Tokenizer}(T_i) \tag{14}$$

The text encoder outputs contextual embeddings:

$$\mathbf{h}_i = f_{\text{BERT}}(\mathbf{x}_i) \tag{15}$$

We extract the CLS token representation as the text feature:

$$\mathbf{t}_i = \mathbf{h}_i^{[CLS]} \in \mathbb{R}^{d_t} \tag{16}$$

where d_t is the hidden dimension of the BERT model.

We use a pretrained ResNet-50 to extract image features. Given an input image I_i, the image embedding is obtained as:

$$\mathbf{v}_i = f_{\text{ResNet}}(I_i) \in \mathbb{R}^{d_v} \tag{17}$$

where d_v is the feature dimension of the ResNet-50 output.

The extracted textual and visual features are projected into a common latent space:

$$\mathbf{t}_i' = W_t \mathbf{t}_i + b_t \in \mathbb{R}^{d_f} \tag{18}$$

$$\mathbf{v}_i' = W_v \mathbf{v}_i + b_v \in \mathbb{R}^{d_f} \tag{19}$$

where d_f is the fusion feature dimension, and W_t, W_v are learnable projection matrices.

The final multimodal feature vector is obtained by concatenation:

$$\mathbf{z}_i = [\mathbf{t}_i'; \mathbf{v}_i'] \in \mathbb{R}^{2d_f} \tag{20}$$

A binary classifier maps the fused representation to a scalar output:

$$\hat{y}_i = \sigma(W_o \mathbf{z}_i + b_o) \tag{21}$$

where W_o and b_o are learnable parameters, and $\sigma(\cdot)$ denotes the sigmoid activation. To classify the meme, a threshold $\tau = 0.5$ is applied:

$$\hat{y} = \begin{cases} 1, & \text{if } y \geq \tau \text{ (Misogyny)}, \\ 0, & \text{otherwise (Non-Misogyny)}. \end{cases}$$

We optimize the model using Binary Cross-Entropy (BCE) loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \Big[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \Big] \tag{22}$$

The model is trained using AdamW optimizer with learning rate η and weight decay λ:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L} + \lambda \theta \tag{23}$$

.

3.4 Majority Voting

The proposed algorithm determines the final classification label for misogyny detection using predictions from three different models (M1, M2, and M3). It employs a majority voting mechanism, where the final label is assigned based on the agreement of at least two out of three used models using Algorithm 1 (Appendix 7).

4 Experimental Settings and Evaluations Metrics

We implemented our model using PyTorch and Hugging Face Transformers, with training conducted on an Nvidia A30 GPU. The model was trained for 12 epochs using an AdamW optimizer with a learning rate of 1e-5 and batch size of 8. For reproducibility, we set the manual seed to 30 and used a dropout rate of 0.3 to prevent overfitting.

4.1 Results and Analysis

Our proposed framework achieved an F1 score of 0.77 for misogyny meme detection in Tamil. For Malayalam, the framework attained an F1 score of 0.84. Tables 2 and 3 present our system's performance compared to other participating systems on the test dataset.

For misogyny meme detection in Tamil languages [2], our system achieved competitive results, ranking 5th among all participants. The best-performing system achieved an F1 score of 0.83682, demonstrating the challenging nature of misogyny meme detection in Tamil languages. In misogyny meme detection in Malayalam languages [3], our system ranked 4th among all participants. While the top system achieved an F1 score of 0.87631, the relatively small performance gap (0.03631) between the first and fourth positions suggests the complexity of the task and the effectiveness of various approaches.

Table 2: Results comparison of top systems for misogyny meme detection in Tamil languages

System	F1	Rank
DLRG_RR	0.83682	1
CUET-NLP_Big_O	0.81716	2
byteSizedLLM	0.80809	3
CUET-823	0.78120	4
Dll5143 (ours)	0.77591	5

Table 3: Results comparison of top systems for misogyny meme detection in Malayalam languages

System	F1	Rank
CUET_Novice	0.87631	1
HerWILL	0.87483	2
One_by_zero	0.86658	3
Dll5143 (ours)	0.84927	4

Detailed performance analysis of all model variants of Tamil 8.1 and Malayalam 8.2 is presented in Appendix 8.

5 Conclusion

In this study, we explored misogyny meme detection challenges in Dravidian-scripted languages through our participation in DravidianLangTech@NAACL 2025. Our experimentation with various models demonstrated that the majority voting mechanism achieved the best performance, with macro F1 scores of 0.77 and 0.84 for Tamil and Malayalam, respectively. The integration of majority mechanisms proved crucial in addressing model biasness in both languages. Despite achieving competitive rankings—5th in Tamil with an F1 score of 0.77 and 4th in Malayalam with an F1 score of 0.84—our analysis revealed persistent challenges as discussed in the section 6.

6 Limitations

Our work advances low-resource language processing by demonstrating how majority voting across models enhances performance. However, the model struggles with unbalanced data, particularly in detecting implicit or body-part-targeted misogyny due to cultural nuances in Tamil (9.1). Despite using robust multilingual models, they may lack script-specific features needed for Dravidian languages, especially for code-mixed or symbolic terms in Malayalam (9.2). Future work could explore specialized pre-training or script-specific

[2] https://drive.google.com/file/d/1yoXmaSYo4ZJT4AaDHzakDvlgnuUem_v6/view

[3] https://drive.google.com/file/d/1zfg30ajYWTCxPuRVaqY-caZafstetCR1/view

models to better distinguish satirical from hateful content, particularly in body-part-related contexts.

Acknowledgments

The authors express their gratitude to the Ministry of Education, the Indian Institute of Information Technology Allahabad, and the Deep Learning Lab at IIITA for providing the resources necessary to complete this work.

References

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, Charmathi Rajkumar, et al. 2024. Overview of shared task on multitask meme classification-unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526.

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *arXiv preprint arXiv:2401.01523*.

Rahul Ponnusamy, Kathiravan Pannerselvam, R Saranya, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, S Bhuvaneswari, Anshid Ka, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in tamil and malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488.

H Shaun, Samyuktaa Sivakumar, R Rohan, Nikilesh Jayaguptha, and Durairaj Thenmozhi. 2024. Quartet@ lt-edi 2024: A svm-resnet50 approach for multitask meme classification-unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 221–226.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of tamilmemes. In *Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation*, pages 7–13.

Ashok Yadav and Vrijendra Singh. 2024. Hatefusion: Harnessing attention-based techniques for enhanced filtering and detection of implicit hate speech. *IEEE Transactions on Computational Social Systems*.

Ashok Yadav and Vrijendra Singh. 2025. Dll5143a@ nlu of devanagari script languages 2025: Detection of hate speech and targets using hierarchical attention network. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 278–288.

7 Appendix A

7.1 Majority Algorithm

Algorithm 1 Majority-Based Final Label Assignment

Require: Predictions from three models: $M1, M2, M3$ where $M1, M2, M3 \in \{0, 1\}$
Ensure: Final classification label $F \in \{0, 1\}$
Initialize: Model predictions $M1, M2, M3$
Compute:
    $count\_0 \leftarrow \sum_{i=1}^{3}(M_i = 0)$ ▷ Count models predicting 0
    $count\_1 \leftarrow \sum_{i=1}^{3}(M_i = 1)$ ▷ Count models predicting 1
if $count\_1 \geq 2$ **then**
    $F \leftarrow 1$ ▷ Assign Misogyny label
else
    $F \leftarrow 0$ ▷ Assign Non-Misogyny label
end if
return F

8 Appendix B

8.1 Misogyny in Tamil Language Results

The performance for misogyny in Tamil using model M1 is shown in Table 4. The performance for misogyny in Tamil using model M2 is presented in Table 5. The performance for misogyny in Tamil using model M3 model is shown in Table 6.

Table 4: Results for misogyny in Tamil using model M1

	prec.	rec.	f1	supp.
0	0.82	0.90	0.86	210
1	0.62	0.47	0.53	74
acc.			0.78	284
macro	0.72	0.68	0.70	284
weighted	0.77	0.78	0.77	284

Table 5: Results for misogyny in Tamil using model M2

	prec.	rec.	f1	supp.
0	0.80	0.87	0.83	210
1	0.52	0.40	0.45	74
acc.			0.75	284
macro	0.66	0.63	0.64	284
weighted	0.73	0.75	0.73	284

Table 6: Results for misogyny in Tamil using model M3

	prec.	rec.	f1	supp.
0	0.83	0.91	0.87	210
1	0.68	0.5	0.57	74
acc.			0.80	284
macro	0.76	0.70	0.72	284
weighted	0.79	0.80	0.79	284

8.2 Misogyny Detection in Malayalam Results

The performance of our proposed framework with different models in Malayalam language is discussed in this section. The performance of model M1 for Malayalam is presented in Table 7.

Table 7: Results for misogyny in Malayalam using model M1

	Prec.	Rec.	F1	Supp.
0	0.91	0.90	0.91	97
1	0.85	0.87	0.86	63
acc.			0.89	160
macro	0.88	0.89	0.88	160
weighted	0.89	0.89	0.89	160

The M2 model in Malayalam achieved an accuracy of 70% on a val set of 160 samples, as shown in Table 8.

Table 8: Results for misogyny in Malayalam using model M2

	Prec.	Rec.	F1	Supp.
0	0.83	0.63	0.72	97
1	0.59	0.80	0.68	63
acc.			0.70	160
macro	0.71	0.72	0.70	160
weighted	0.74	0.70	0.70	160

The M3 model in Malayalam achieved an accuracy of 85% on val set of 160 samples, as shown in Table 9.

Table 9: Results for misogyny in Malayalam using model M3

	Prec.	Rec.	F1	Supp.
0	0.83	0.93	0.88	97
1	0.88	0.71	0.78	63
acc.			0.85	160
macro	0.85	0.82	0.83	160
weighted	0.85	0.85	0.84	160

9 Appendix C (Error Analysis)

In the error analysis, challenges like the diglossic nature of Tamil, Sandhi (joining of words), and code-mixing add significant complexity to error detection. Understanding these linguistic features is essential for an overall analysis. Tamil is highly diglossic, meaning it has two forms of the same language, used for different purposes, and classical elements and complex morphology make its further interpretation difficult. We can take the case of Sandhi Rules, which means rules for joining two words. In Malayalam, the joining is straightforward without much alteration of the internal letters. In Tamil, words are merged smoothly, but there are letter changes. So we can highlight that it will be easier for tokenizers to handle the case of Malayalam, and as a result, we obtain a better interpretation of the context, whereas the same in the case of Tamil is difficult because it poses a challenge for transformers to split and interpret accurately. Furthermore, there are classical elements and alternatives to Tamil that make interpretation difficult; for example, most memes contain modern Tamil and might miss references to classical alternatives of words that mean the same.

Furthermore, in the Tamil dataset, most of the memes were of a code-mixed nature. And it is a known challenge in hate speech identification in code-mixed languages such as Tamil-English or Malayalam-English. It is much more challenging because of inconsistent language patterns that include vocabulary and grammar shifts that make context interpretation challenging here, as there is a switch between scripts, making tokenization and understanding difficult, and further, like we highlighted, the vocabulary confusion and grammar differences. Whereas most of the Malayalam memes were in Malayalam script only, with limited code-mixed content, explaining the performance gap between Tamil and Malayalam. We also observed that the Tamil dataset was more skewed compared to the Malayalam dataset. In the Tamil

dataset, the percentage of misogynistic memes is only 25% while that of the Malayalam data set is 40%, so the model is biased towards the majority class. Such class imbalance makes the model biased towards the majority class. Some of the misclassified examples are discussed in subsequent subsections in detail.

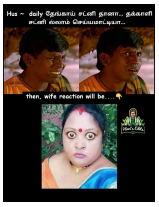
9.1 Appendix C -I

The model exhibited specific challenges in classifying memes containing symbolic language, named entities, and code-mixed expressions, particularly in understanding nuanced cultural and script-specific references in Tamil. Table 10 shows representative examples that illustrate these limitations.

The meme (ID: 1097) contains text that includes words like 'dress' and 'floor' alongside Tamil words, scripted in English. This is a classic example of code-mixing, demonstrating how non-standard vocabulary may not be recognized by the pretrained model, making comprehension more challenging and adding complexity. As a result, M1 misclassifies it, failing to detect the sarcastic tone and non-standard vocabulary, which imply 'women are short.' However, M2 and M3 effectively detect misogyny, highlighting the advantage of using multiple models to capture such nuances. The meme (ID: 1163) is flagged as misogynistic by both models M2 and M3, while model M1 views the content as non-misogynistic, contrary to the original misogynistic label, as it captures the reinforcing negative stereotypes about women as manipulative and deceitful. It uses the notion of "playing" with emotions, which can be seen as trivializing the sincerity of relationships and portraying women negatively. The meme (ID: 1329) with the original label as misogynistic is captured wrongly by M2 due to different sensitivity levels and failure to capture the negative stereotypes properly, especially when playing into stereotypes of intrafamilial female conflict, and hence couldn't capture the negative views about the behavior of women within family dynamics.

The meme (ID: 1431) reflects domestic stereotypes and cultural critique. It primarily contains Tamil text with some English, representing the wife's reaction. The text translates to: 'You only make coconut chutney every day. Don't you ever make tomato chutney?' This critique of a woman's cooking habits reflects a common domestic stereotype in Tamil culture, where women are expected to manage household chores. Such cultural nuances influence prediction, as recognizing implicit domestic criticism and stereotypes requires models to understand societal expectations and gender roles prevalent in Tamil culture. Models M1 and M3 identified the meme as misogynistic based on these implicit stereotypes and cultural patterns. However, Model M2 misinterpreted it, viewing the husband's complaint as marital humor rather than a targeted critique of women.

Table 10: Error Analysis of Samples for Misogyny in Tamil

Image ID	M1	M2	M3	Actual Label	Image
1097	0	1	1	1	
1163	0	1	1	1	
1329	1	0	1	1	
1431	1	0	1	1	
1340	1	0	1	1	
1643	0	1	1	0	
1639	0	1	1	0	

The meme (ID: 1340) is labeled misogynistic for objectifying or inappropriately commenting on a woman's attire. This highlights differences in how models interpret satire and humor, with M1 and

M3 recognizing misogyny, while M2 fails to do so.

The meme (ID: 1643) is labeled non-misogynistic in the original dataset, but M3 and M2 interpret certain textual or visual elements as misogynistic due to cultural context or direct translations linked to gender stereotypes. In contrast, M1 does not classify it as misogynistic. The meme (ID: 1639), originally labeled as non-misogynistic, presents a challenging scenario for model interpretation. While Models M2 and M3 detect misogyny based on visual or textual signals, Model M1 does not. This suggests that M1 better grasps context or nuances like sarcasm, which the other models overlook. In addressing the challenges posed by the detection of misogynistic content in visual media, the utilization of varying capabilities of these models to interpret nuances and contextual cues—where M1 sometimes outperforms M2 and M3 in recognizing sarcasm and cultural contexts—illustrates the diversity in model training and perspective. This diversity is instrumental in capturing a wider range of interpretations that might be missed by a single model. Even though some inaccuracies still persist in cases of majority voting, this pragmatic blend of accounting insights from multiple models ensures a more consistent and accurate interpretation in the majority of cases.

9.2 Appendix C -II

To evaluate our model's robustness in distinguishing misogyny and non-misogyny, we conducted an error analysis on misclassified instances. This analysis provided insights into common misclassification patterns. Table 11 presents examples of these errors along with interpretations for each case.

The meme (ID: 954) was originally labeled as non-misogynistic. Models M2 and M3 correctly identified that the phrase conveys admiration without any negative or objectifying language toward women. However, Model M1 misinterpreted the context, associating keywords like 'children' or 'young fellows' with biased content. Majority voting helped maintain stability, ensuring the final decision aligned with the correct label.

The meme (ID: 239) and original label as non-misogynistic, even though this was just a general discussion, model M3 predicted it as misogynistic because of the presence of both the gender and model's lack of understanding of context in deeper levels, flagging terms like "he" or "childhood" as signals for gender discussions even when the con-

Table 11: Error Analysis of Samples for Misogyny in Malayalam

Image ID	M1	M2	M3	Actual Label	Image
954	1	0	0	0	
239	0	0	1	0	
545	1	1	0	1	
725	0	1	1	1	
112	0	1	0	1	
649	1	1	0	1	
168	0	0	1	0	
317	0	0	0	0	

tent doesn't support such an interpretation. The absence of objectifying or specific gender references should have indicated a non-related context, while models M1 and M2 demonstrate a good contextual understanding. As a result, the majority voting predicts the correct label; this can again be observed with the meme (ID: 545), which translates to "Asking the neighbor's sister for a game," which directly pertains to objectification and inappropriate propositions. Again, Model M3 fails to understand the context of intent, while Models M1 and M2 correctly identify the content as related by recognizing the objectifying request towards women.

Moving to the meme (ID: 725), which is labeled as misogynistic and makes remarks about women's "backs" and has an image of a woman and the phrase "Back, That's all". Here, model M1 failed to map the woman's image and its phrase and recognize the context in which the word "back" is used, while models M2 and M3 were effective in detecting objectifying language even when it pertains to specific body parts, and so the majority voted.

The meme (ID: 112) labeled as misogynistic is predicted the opposite by majority voting, and models M1 and M3 failed to map the remarks "flower," which actually made remarks about women's body parts. Both models failed to detect the explicit objectifying content, possibly due to focusing on specific keywords without fully understanding the context or the nature of the remarks. While Model M2 was effective in mapping and recognizing the indirect remarks about women's body parts, the suggests we need to further work on these models to make them more robust and make these model's aware to map these indirect references correctly.

The meme (ID: 649) is marked as misogynistic, with the translated keyword "There's an aunty like this everywhere, to stir up the crowd," where model M1 and M2 were correctly able to catch the explicit objectifying language, whereas we can say that model M3 requires better contextual sensitivity to identify nuanced objectifying or critical statements about gender dynamics.

The meme (ID: 168) depicts a lighthearted conversation: 'BF: When are we getting married? GF: When the movie releases. And that movie is not getting released anytime soon.' It is labeled as non-misogynistic as it simply portrays a humorous situation. However, Model M3 incorrectly predicts it as misogynistic, associating it with negative stereotypes related to women's reliability or commitment. In contrast, the other models correctly identify it as unrelated to misogyny. Detecting misogynistic content in Malayalam visual media presents significant challenges, particularly in interpreting nuanced cultural contexts, implicit biases, and subtle forms of objectification. While leveraging the strengths of different models and majority voting helps mitigate individual model biases, it occasionally propagates errors when multiple models misclassify similar content. However, majority voting generally improves performance in ambiguous cases. For example, in the case of meme (ID: 317), the models remain accurate despite the presence of women in the image, as the context is political rather than misogynistic.

Nevertheless, the complexity of Malayalam's linguistic and societal nuances, subtle misogyny, and indirect references remains a challenge. Addressing these issues requires integrating culturally specific and contextually rich datasets, enhancing models' ability to recognize implicit biases, and incorporating advanced techniques such as weighted ensemble methods and fine-grained contextual embeddings in future research.

KEC_AI_VSS_run2@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media

Kogilavani Shanmugavadivel[1],**Malliga Subramanian**[2],
Sathiya Seelan S[1], **Suresh Babu K**[1], **Vasikaran S**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{sathiyaseelans.22,sureshbabuk.22,vasikarans.22aid}@kongu.edu

Abstract

The increasing instances of abusive language against women on social media platforms have brought to the fore the need for effective content moderation systems, especially in low-resource languages like Tamil and Malayalam. This paper addresses the challenge of detecting gender-based abuse in YouTube comments using annotated datasets in these languages. Comments are classified into abusive and non-abusive categories. We applied the following machine learning algorithms, namely Random Forest, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting and AdaBoost for classification. Micro F1 score of 0.95 was achieved by SVM for Tamil and 0.72 by Random Forest for Malayalam. Our system participated in the shared task on abusive comment detection,out of 160 teams achieving the rank of 13th for Malayalam and rank 34 for Tamil, and both indicate both the challenges and potential of our approach in low-resource language processing. Our findings have highlighted the significance of tailored approaches to language-specific abuse detection.

1 Introduction

Social media has revolutionized communication, interaction, information sharing, and expression. However, its misuse has led to serious issues, including gender-based abuse targeting women. Such abusive language is mostly derogatory and threatening, which reflects societal biases and has serious psychological, social, and professional consequences for the victims.Tamil and Malayalam are low-resource languages that lack robust automated systems to address this challenge. Identifying abusive content in these languages is important for effective moderation. This paper focuses on the detection of gender-based abuse in Tamil and Malayalam YouTube comments.This can help in creating safer and more inclusive online environments.

Detecting abusive content in low-resource languages is challenging due to the scarcity of annotated datasets and the complexity of linguistic nuances. The implicit bias, coded expressions, and slangs used in abusive language make it difficult for the automated system to detect Tamil and Malayalam. We approach these challenges by curating annotated datasets of YouTube comments in Tamil and Malayalam. Each comment is labeled as either abusive or non-abusive, and examples reflect explicit and implicit abuse. Developing accurate classification models for these datasets is important for enhancing content moderation systems.

We, therefore, use machine learning algorithms such as Random Forest, SVM, KNN, Gradient Boosting, and AdaBoost on the comments dataset to classify them as either abusive or non-abusive. The experiments show that SVM gives the highest micro F1 score at 0.95 for Tamil, but Random Forest gives the best score of 0.72 for Malayalam. These findings reflect the efficiency of language-specific models in detecting abusive content. The empirical findings from this research study will help to solve the imperative case of gender-based abuse issues in social media and make cyberspace safer for women.

2 Literature Survey

The Multimodal Tamil Hate (MATH) dataset has been introduced for detecting hate speech in Tamil across text, audio, and video modalities Mohan et al. (2023). A combination of BERT for text, TimeSformer for video, and Wav2vec2 for audio was used, achieving 81.82% accuracy with a multimodal fusion approach. Tamil abusive comment classification has been improved through multilingual transformers and data augmentation, leading to a 15-unit increase in the macro F1-score using the MURIL model Sheik et al. (2023). Additionally, Tamil and code-mixed Tamil-English datasets for

abusive comment detection on YouTube have been created, showing classical models as more effective due to limited data Chakravarthi et al. (2023).

A transformer-based approach has been proposed for detecting abusive content in 13 Indic code-mixed languages, outperforming classical models. The combination of XLM-RoBERTa with BiGRU and emoji embeddings achieved an F1 score of 0.88 and an AUC of 0.94 Bansal et al. (2022). Multilingual embeddings like IndicBERT and MuRIL have been utilized for Tamil and Telugu, demonstrating superior performance over classical models on YouTube datasets Vegupatti et al. (2023).

A toxic comment detection system for Assamese has been developed using SVM with TF-IDF, achieving 94% accuracy and F1-score Dutta et al. (2024). An overview of a shared task on detecting abusive and hope speech in Tamil and Tamil-English social media comments has been provided, employing various machine learning and deep learning methods Priyadharshini et al. (2022).

Studies from the Third Workshop on Speech and Language Technologies for Dravidian Languages have presented insights into abusive comment detection Priyadharshini et al.. Logistic regression with embeddings has been explored for Tamil and Telugu datasets Bala and Krishnamurthy (2023). A study has achieved 99% accuracy in abusive comment detection for code-mixed Tamil-English text Pannerselvam et al. (2023). Machine learning, deep learning, and BERT have been employed for Tamil, achieving notable rankings in a shared task Shanmugavadivel et al. (2023).

3 Task Description

This task concentrated on finding abusive comments toward women in the language of Tamil and Malayalam from social media sites, such as YouTube. We annotated the data carefully with proper labels in both abusive and non-abusive categories to have a high-quality set of labels for supervised learning. We utilized several machine learning models, namely SVM, Random Forest, KNN, Gradient Boosting, and AdaBoost to classify comments against this challenge. SVM achieved the highest micro F1 score of 0.95 for Tamil, while Random Forest performed best for Malayalam with a score of 0.72. The models were trained and fine-tuned using various feature extraction techniques, including TF-IDF and word embeddings, to capture linguistic nuances.Our system participated in the shared task Rajiakodi et al. (2025) on abusive comment detection, ranking 13th for Malayalam and 34th for Tamil out of 160 teams. This therefore shows the problems of abusive language detection in low-resource languages and the need to develop robust language-specific content moderation systems.

4 Dataset Description

The Tamil dataset has a total of 2,790 records split between the classes evenly. Therefore, it has 1,366 abusive comments and 1,402 non-abusive comments, which makes the dataset balanced for unbiased training and model testing. A balanced dataset means that none of the models gets biased to prefer any of the classes, one being abusive and the other being not.

The Malayalam dataset has 2,933 records with a little imbalance in its distribution. This dataset contains 1,531 abusive comments and 1,402 non-abusive comments; this is true to real life, where most of the instances of abusive language are dominant in certain contexts.This dataset is a bit more challenging as it also is imbalanced and includes the complexity of the Malayalam language.

Language	Abusive(N)	Non Abusive
Malayalam	1531	1402
Tamil	1366	1424

Table 1: Dataset Description

5 Methodology

The methodology for abusive comment detection in Tamil and Malayalam consists of data preprocessing, feature extraction, and model development. Each step ensures efficient processing by transforming raw data into meaningful representations using machine learning-based classification.

5.1 Data Preprocessing

Data preprocessing involves cleaning raw YouTube comments for classification. This includes removing special characters, emojis, punctuation, and URLs. Tokenization splits each comment into words or phrases for analysis, followed by stopword removal to enhance computational efficiency. Next, text normalization standardizes spelling and slang variations in Tamil and Malayalam. Labels are encoded into numbers for machine learning

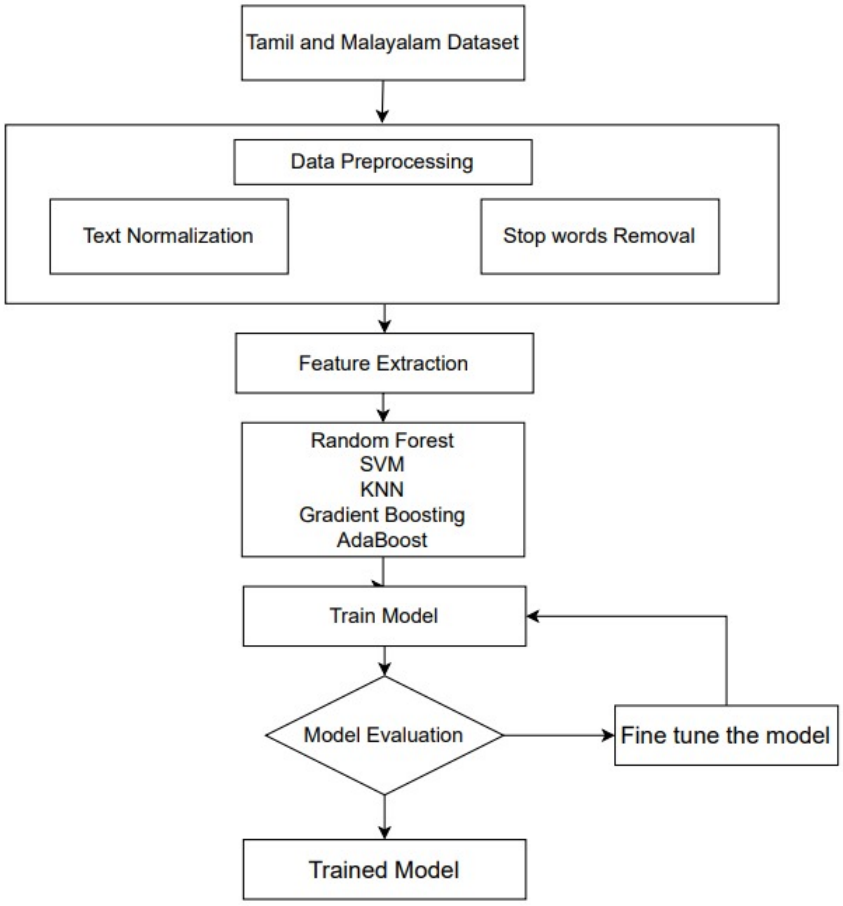


Figure 1: Proposed System Workflow

compatibility. To address class imbalance, methods like over-sampling the majority class and under-sampling the minority class are applied, ensuring better model performance by preventing bias toward one class.

5.2 Feature Extraction

After preprocessing, features are extracted using methods like TF-IDF (Term Frequency-Inverse Document Frequency), which weighs words based on their frequency in abusive comments while down-weighting common terms. Additionally, word embeddings like Word2Vec, FastText, and BERT capture contextual meaning, enabling the model to understand relationships between words. These features are crucial for enhancing model precision in distinguishing abusive from non-abusive language.

5.3 Model Development

The final step involves selecting, training, and testing machine learning models for abusive comment classification. The dataset is split into training, validation, and test sets for effective generalization. Several algorithms, including SVM, Random Forest, KNN, Gradient Boosting, and AdaBoost, are fine-tuned for optimal performance. Evaluation metrics such as accuracy, precision, recall, and micro F1 score are calculated. SVM performs best for Tamil with a micro F1 score of 0.95, while Random Forest achieves the highest micro F1 score of 0.72 for Malayalam, indicating its effectiveness in handling linguistic variations in abusive comments. The workflow diagram in Figure 1 shows the entire process, from preprocessing to model evaluation.

6 Experimental Analysis

To check the performance of the models, Macro-F1 score is employed that is widely used in classification problems, especially on imbalanced datasets. The experiments are carried out on text data for both the Tamil and Malayalam languages. Accordingly, the corresponding performances are explained in the subsections below.

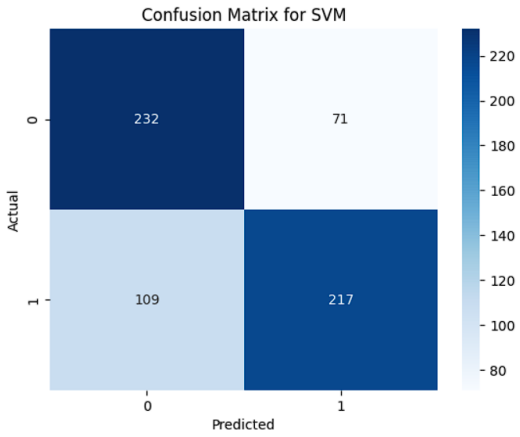


Figure 2: Confusion Matrix of Tamil Data

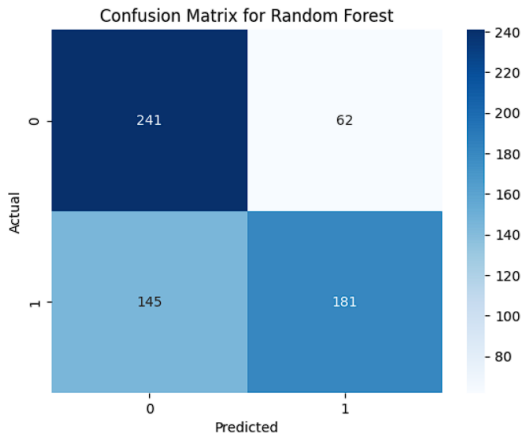


Figure 3: Confusion Matrix of Malayalam Data

6.1 Tamil

For this, the model performance was analyzed with accuracy, precision, recall, and micro F1 score metrics to compare their ability in classifying the comments as abusive or non-abusive. For the Tamil dataset, Support Vector Machine has been found out to be performing better than others with a micro F1 score of 0.95, which is really a high score. This is due to the reason that SVM handles the linguistic variations in Tamil such as colloquial terms and context-dependent variations. Figure 2 Depicts the best-performing model's confusion matrix on Tamil data. The best model was able to strongly capture the explicit and implicit abusive language patterns and, therefore could be relied on for Tamil text classification.

6.2 Malayalam

For the Malayalam dataset, Random Forest was the best model with a micro F1 score of 0.72. Although the score is lower than that of Tamil, it shows that Random Forest generalizes well despite the complexity of Malayalam grammar and vocabulary. The model was able to capture patterns of abuse, including slang and implicit bias in Malayalam comments. This performance depicts robustness and applicability for moderate abusive content in the Malayalam language effectively. Figure 3: Confusion matrix of the best performing model on Malayalam data.

7 Limitations

This study has several limitations. First, the dataset size is limited, which may affect the generalizability of the models. Second, the imbalance in the Malayalam dataset could lead to biased predictions. Third, the approach relies on traditional machine learning models, which might not capture complex linguistic patterns effectively. Lastly, the absence of deep learning techniques limits the potential for higher accuracy.

8 Conclusion

The research throws light on the increasing phenomenon of gender-based abuse on social media and the growing need for automatic content moderation, especially in low-resource languages such as Tamil and Malayalam. A classification system has been developed for effectively detecting abusive comments by applying SVM, Random Forest, KNN, Gradient Boosting, and AdaBoost machine learning models. Our results indicate that SVM obtained the best micro F1 score of 0.95 for Tamil, while the best-performing model for Malayalam was Random Forest with a score of 0.72. Our system has also participated in the shared task on abusive comment detection and ranked 13th in Malayalam and 34th in Tamil out of 160 teams. Deep models and contextualized embeddings are much more beneficial for the advancement of this problem, and future work will be directed towards enlarging datasets and fine-tuning classification techniques to make them more accurate and generalizable for real-world applications. The code for this shared task can be accessed at Github

References

Abhinaba Bala and Parameswari Krishnamurthy. 2023. AbhiPaw@ DravidianLangTech: Abusive comment detection in Tamil and Telugu using logistic regression. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian*

Languages, pages 231–234, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Vibhuti Bansal, Mrinal Tyagi, Rajesh Sharma, Vedika Gupta, and Qin Xin. 2022. A transformer based approach for abuse detection in code mixed indic languages. *ACM transactions on Asian and low-resource language information processing*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Surajit Dutta, Mandira Neog, and Nomi Baruah. 2024. Assamese toxic comment detection on social media using machine learning methods. In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, pages 1–8. IEEE.

Jayanth Mohan, Spandana Reddy Mekapati, and Bharathi Raja Chakravarthi. 2023. A multimodal approach for hate and offensive content detection in tamil: From corpus creation to model development. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Kathiravan Pannerselvam, Saranya Rajiakodi, Rahul Ponnusamy, and Sajeetha Thavareesan. 2023. CSS-CUTN@DravidianLangTech:abusive comments detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 306–312, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and booktitle = Kumaresan, Prasanna Kumar". Findings of the shared task on Abusive Comment Detection in Tamil and Telugu.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Shri Durga R, Srigha S, Sree Harene J S, and Yasvanth Bala P. 2023. KEC_AI_NLP@DravidianLangTech: Abusive comment detection in Tamil language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 293–299, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Reshma Sheik, Raghavan Balanathan, and Jaya Nirmala S. 2023. Mitigating abusive comment detection in Tamil text: A data augmentation approach with transformer model. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 460–465, Goa University, Goa, India. NLP Association of India (NLPAI).

Mani Vegupatti, Prasanna Kumar Kumaresan, Swetha Valli, Kishore Kumar Ponnusamy, Ruba Priyadharshini, and Sajeetha Thavaresan. 2023. Abusive social media comments detection for tamil and telugu. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 174–187. Springer.

The_Deathly_Hallows@DravidianLangTech 2025: AI Content Detection in Dravidian Languages

Kogilavani Shanmugavadivel[1] **, Malliga Subramanian**[2]**,**
Vasantharan K[1]**, Prethish G A**[1]**, Vijayakumaran S**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{vasantharank.ncc, prethish0409, vijayakumaran2k3}@gmail.com

Abstract

The DravidianLangTech@NAACL 2025 shared task focused on Detecting AI-generated Product Reviews in Dravidian Languages, aiming to address the challenge of distinguishing AI-generated content from human-written reviews in Tamil and Malayalam. As AI generated text becomes more prevalent, ensuring the authenticity of online product reviews is crucial for maintaining consumer trust and preventing misinformation. In this study, we explore various feature extraction techniques, including TF-IDF, Count Vectorizer, and transformer-based embeddings such as BERT-Base-Multilingual-Cased and XLM-RoBERTa-Large, to build a robust classification model. Our approach achieved F1-scores of 0.9298 for Tamil and 0.8797 for Malayalam, ranking 8th in Tamil and 11th in Malayalam among all participants. The results highlight the effectiveness of transformer-based embeddings in differentiating AI-generated and human-written content. This research contributes to the growing body of work on AI-generated content detection, particularly in underrepresented Dravidian languages, and provides insights into the challenges unique to these languages.

.

1 Introduction

With the rapid advancement of natural language generation models, AI-generated text has become increasingly prevalent across various domains, including online product reviews. While these models enhance automation and efficiency, they also raise concerns regarding the authenticity and reliability of online content. AI-generated reviews have the potential to influence consumer decisions, mislead potential buyers, and distort market perceptions. This issue is particularly critical in low-resource languages such as Tamil and Malayalam, where limited research has been conducted on detecting AI-generated content Premjith et al. 2025.

To address this challenge, the Shared Task on Detecting AI-generated Product Reviews in Dravidian Languages was organized, providing a dataset consisting of both human-written and AI-generated reviews in Tamil and Malayalam. Participants were tasked with developing models to accurately classify these reviews while considering the unique linguistic complexities of Dravidian languages. The evaluation metric used in this task was the F1-score, ensuring a balanced and robust assessment of model performance.

In this work, we present our methodology and findings in tackling this problem. We explored both traditional feature extraction methods, such as TF-IDF and Count Vectorizer, and transfer learning models, including BERT-Base-Multilingual-Cased and XLM-RoBERTa-Large. Through extensive experimentation and analysis, we highlight the challenges involved in distinguishing AI-generated reviews from human-authored ones and assess the effectiveness of various feature extraction techniques and classification models. Our results provide insights into the applicability of transformer-based embeddings for AI-generated content detection in underrepresented Dravidian languages.

2 Literature Review

The study conducted by Wu et al. 2023 examined the role of ChatGPT in credit default prediction by comparing AI-generated and human-generated loan assessments. Their findings indicated that ChatGPT-generated insights contributed to improved predictive accuracy, underscoring the model's potential in financial decision-making. Agrawal et al. 2019 investigated the impact of AI on human labor, particularly differentiating between prediction and judgment tasks. The study demonstrated that AI significantly reduced predic-

tion costs, leading to increased variance in outcomes and altering the perceived value of human judgment in decision-making processes. Molina and Sundar 2024 explored the factors influencing user trust in AI-driven content moderation. Their research found that individuals with lower trust in human moderators were more inclined to favor AI-based moderation systems, providing insights into the evolving trust dynamics between automation and human decision-making.

Cao et al. 2023 conducted a comprehensive survey on AI-generated content (AIGC), tracing its evolution from early generative adversarial networks (GANs) to advanced models such as ChatGPT and DALL-E-2. The study outlined significant advancements, identified key challenges, and discussed future directions for generative AI applications. Singhal and Bedi 2024 presented a transformer-based approach for Tamil code-mixed sentiment analysis, specifically applied to hate speech detection. Their ensemble model achieved the highest ranking at LT-EDI 2024, demonstrating the effectiveness of RoBERTa-based architectures in multilingual and code-mixed settings. Devananthan and Nair 2023 examined multilingual sentiment analysis on Indian Twitter, evaluating a range of machine learning and deep learning models. Their research contributed to the development of a robust framework capable of processing diverse linguistic content efficiently. Kumaresan et al. 2022 focused on hope speech detection in code-mixed Tamil, English, and Malayalam. The study employed transformer-based models that achieved competitive F1 scores, highlighting their effectiveness in sentiment classification and content moderation tasks. Li et al. 2021 proposed a cross-lingual named entity recognition (NER) approach utilizing XLM-RoBERTa in conjunction with parallel corpora. Their approach enhanced entity alignment without the need for direct translation, surpassing the performance of unsupervised methods across multiple languages. Gaikwad et al. 2023 conducted a comparative analysis of multilingual sentiment analysis models. Their findings revealed that XLM-RoBERTa attained the highest accuracy (78.37%), demonstrating its adaptability and efficacy across various linguistic contexts. Raja et al. 2023 focused on fake news detection in Malayalam by optimizing an XLM-RoBERTa model. Their model achieved a macro-averaged F-score of 87%, securing the second rank in the DravidianLangTech competition, further reinforcing the effectiveness of transformer-based models in tackling misinformation detection tasks.

3 Dataset Description

The dataset for this task consists of AI-generated and human-written product reviews in two Dravidian languages: Malayalam and Tamil. The dataset was created to support the development of models capable of distinguishing between machine-generated and human-authored content in online reviews .

The Tamil dataset comprises 808 samples, with 405 AI-generated reviews and 403 human-written reviews. The Malayalam dataset contains 800 samples, evenly split between AI-generated and human-written reviews. Each sample is a textual review, and the datasets provide a balanced distribution to ensure fair model evaluation.

Dataset	AI-Generated	Human
Tamil	405	403
Malayalam	400	400

Table 1: Distribution of AI-generated and human-written reviews in the dataset.

4 Task Description

To tackle the increasing difficulty of identifying AI-generated material in online reviews, the Shared Task on Detecting AI-Generated Product Reviews in Dravidian Languages (DravidianLangTech@NAACL 2025) was created. As AI models get more complex, it is essential to differentiate between evaluations written by humans and those written by robots in order to preserve credibility and confidence in online markets. In Malayalam and Tamil, participants will create and assess models that categorise product reviews as either human-written or AI-generated. For training and testing, the dataset will be made available in an organised manner. Participants can download the dataset and submit their models for review on CodaLab.

The F1-score, a commonly used statistic for classification tasks in NLP, will be used to evaluate the model's performance. Participants from a variety of academic disciplines are invited to participate in this shared challenge, which aims to improve the recognition of AI-generated material in low-resource languages.

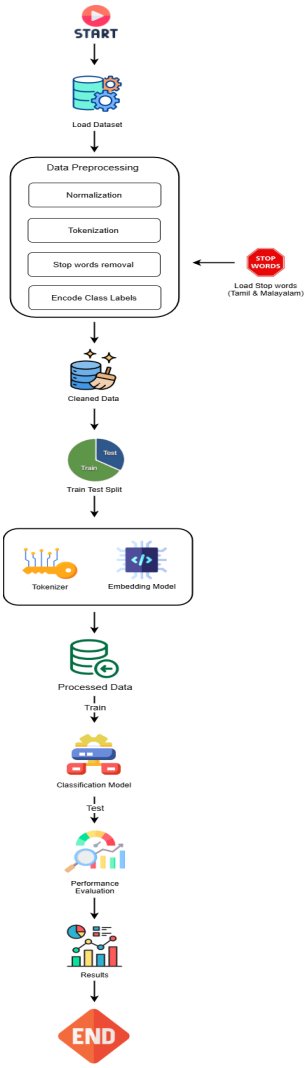


Figure 1: Proposed Model Workflow

5 Methodology

5.1 Preprocessing Dataset

The collection includes both human-written and AI-generated product reviews in Malayalam and Tamil. An ID, a textual review (transcript), and a class label specifying whether the review was prepared by a person or by artificial intelligence make up each instance.

Because Malayalam stopwords were not available in Advertools, stopwords for Malayalam were retrieved from a publicly accessible Git repository, whereas stopwords for Tamil were eliminated using the Advertools library. Tokenisation, stopword elimination, and text normalisation were among the preparation procedures. The *Indic NLP Library* was used to normalise the Tamil text, and the *indic-tokenize* package was used to tokenise it. After preprocessing, the text was saved for further use.

5.2 Feature Extraction

To obtain meaningful text representations, three different feature extraction techniques were used:

1. **TF-IDF:** Term Frequency-Inverse Document Frequency (TF-IDF) was used to convert text into numerical vectors.

2. **Count Vectorizer:** This method created a sparse representation of word occurrences in the corpus.

3. **Transformer-Based Embeddings:** Two pre-trained transformer models, BERT-Base-Multilingual-Cased and XLM-RoBERTa-Large, were used to extract contextual embeddings. The models were loaded using the Hugging Face transformers library.

For transformer-based embeddings, tokenization was performed using the corresponding model's tokenizer. The processed text was converted into tokenized sequences with a maximum length of 512 tokens. The embeddings were extracted by taking the mean of the last hidden state outputs of the model.

5.3 Classification Model

A deep learning-based classification model was developed to distinguish AI-generated reviews from human-written reviews. The model architecture included:

- A fully connected dense layer with 256 neurons and ReLU activation.
- Batch normalization and dropout (0.5) to prevent overfitting.
- Another dense layer with 128 neurons and ReLU activation, followed by batch normalization and dropout (0.5).
- A final dense output layer with softmax activation for classification.

The model was trained using the Adam optimizer with categorical cross-entropy loss. The dataset was split into 80% training and 20% testing using the *train_test_split* function from Scikit-learn.

6 Limitations

This study faces two primary limitations. First, the high computational cost of transformer-based models like XLM-RoBERTa-Large and BERT-Base-Multilingual-Cased makes them resource-intensive,

Models Used	Tamil	Malayalam
BERT-Base-Multilingual-Cased	94%	94%
TF-IDF	81%	76%
Count Vectorizer	83%	76%
XLM-RoBERTa-Large	96%	94%

Table 2: Performance of Different Models in Tamil and Malayalam in text

limiting their accessibility to researchers with constrained computational resources. Second, the dataset size is relatively small, with only 808 reviews in tamil and 800 in malayalam, which may affect the model's ability to generalize across different domains such as social media or news articles. Expanding the dataset and optimizing models for efficiency would enhance the applicability of AI content detection in Dravidian languages.

7 Performance Evaluation

Using the Accuracy, we assessed several text representation methods and classification models. Conventional techniques such as Count Vectorizer and TF-IDF shown difficulties in capturing contextual semantics, achieving accuracy of 76% for Malayalam and 81% and 83% for Tamil.

Transformer-based models performed noticeably better than conventional methods. While XLM-RoBERTa-Large performed the best, achieving 96% for Tamil and 94% for Malayalam, BERT-Base-Multilingual-Cased had an accuracy of 94% for both languages. These findings demonstrate how well deep contextual embeddings work to differentiate between writing produced by AI and text authored by humans.

8 Conclusion

This study explored various feature extraction techniques and classification models to distinguish AI-generated and human-written product reviews in Tamil and Malayalam. Traditional methods such as TF-IDF and Count Vectorizer were found to be less effective due to their inability to capture contextual semantics. Transformer-based models, particularly XLM-RoBERTa-Large, provided the highest accuracy, demonstrating the effectiveness of deep contextual embeddings. The results emphasize the importance of using pre-trained multilingual models for low-resource languages. By leveraging transformer-based architectures, we achieved an Accuracy of 96 for Tamil and 94 for Malayalam, outperforming traditional statistical approaches. Future research can explore fine-tuning transformer models on larger domain-specific datasets and incorporating additional linguistic features to further enhance classification accuracy. Additionally, integrating explainable AI techniques could provide insights into model decision-making, making AI-generated content detection more interpretable and trustworthy. The source code for our approach is available at https://github.com/vasantharan/Detecting_AI_generated_product_reviews_in_Dravidian_languages.

References

Ajay Agrawal, Joshua S Gans, and Avi Goldfarb. 2019. Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, 47:1–6.

Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*.

AG Devananthan and Lekshmi S Nair. 2023. Exploring multilingual indian twitter sentiment analysis: A comparative study. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–8. IEEE.

Arya Gaikwad, Pranav Belhekar, and Vinayak Kottawar. 2023. Advancing multilingual sentiment understanding with xgboost, svm, and xlm-roberta. In *International Conference on Data Science, Machine Learning and Applications*, pages 990–1000. Springer.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Bing Li, Yujie He, and Wenjin Xu. 2021. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *arXiv preprint arXiv:2101.11112*.

Maria D Molina and S Shyam Sundar. 2024. Does distrust in humans predict greater trust in ai? role

of individual differences in user responses to content moderation. *New Media & Society*, 26(6):3638–3656.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Eduri Raja, Badal Soni, and Sami Kumar Borgohain. 2023. nlpt malayalm@ dravidianlangtech: Fake news detection in malayalam using optimized xlm-roberta model. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 186–191.

Kriti Singhal and Jatin Bedi. 2024. Transformers@ dravidianlangtech-eacl2024: Sentiment analysis of code-mixed tamil using roberta. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 151–155.

Zongxiao Wu, Yizhe Dong, Yaoyiran Li, and Baofeng Shi. 2023. Unleashing the power of text for credit default prediction: Comparing human-generated and ai-generated texts. *Available at SSRN 4601317*.

SSN_MMHS@DravidianLangTech 2025: A Dual Transformer Approach for Multimodal Hate Speech Detection in Dravidian Languages

Jahnavi Murali and **Rajalakshmi Sivanaiah**
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai, India
jahnavi2110854@ssn.edu.in, rajalakshmis@ssn.edu.in

Abstract

The proliferation of the Internet and social media platforms has resulted in an alarming increase in online hate speech, negatively affecting individuals and communities worldwide. While most research focuses on text-based detection in English, there is an increasing demand for multilingual and multimodal approaches to address hate speech more effectively. This paper presents a methodology for multiclass hate speech classification in low-resource Indian languages namely, Malayalam, Telugu, and Tamil, as part of the shared task at DravidianLangTech 2025. Our proposed approach employs a dual transformer-based framework that integrates audio and text modalities, facilitating cross-modal learning to enhance detection capabilities[1]. Our model achieved macro-F1 scores of 0.348, 0.1631, and 0.1271 in the Malayalam, Telugu, and Tamil subtasks respectively. Although the framework's performance is modest, it provides valuable insights into the complexities of multimodal hate speech detection in low-resource settings and highlights areas for future improvement, including data augmentation and alternate fusion and feature extraction techniques.

1 Introduction

The rise of social media has amplified the spread of hate speech, making it a pervasive and pressing issue. Online hate harms both victims and observers, often leading to issues like depression, isolation, social anxiety, and loss of confidence (Walther, 2022). To address this problem and ensure a safer online space, researchers have focused extensively on developing hate speech detection methods for text-based content, while detection in audio data remains underexplored, presenting unique challenges and opportunities (Bhesra et al., 2024). Furthermore, most studies in this domain are primarily limited to English, underscoring the critical need for robust multilingual and multimodal hate speech detection systems(Chhabra and Vishwakarma, 2023; Nandi et al., 2024) in low-resource monolingual and code-mixed languages (Premjith et al., 2024a).

This work aims to advance the development of efficient multimodal fine-grained hate speech detection systems for low-resource Dravidian languages by utilizing intermediate fusion techniques to enhance cross-modal learning. The paper is organized as follows: Previous research on hate speech detection in multimodal and multilingual settings are discussed in Section 2. The datasets used and the proposed system along with the architecture diagram are outlined in Sections 3 and 4 respectively. Results obtained are presented in Section 5, and the paper concludes with intent and direction for future work.

2 Related Works

Detecting hate speech in low-resource and code-mixed languages presents unique challenges due to linguistic diversity and the lack of annotated data sets. Several studies have explored different approaches to tackle this issue. Premjith et al. (2024b) reported on the Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) shared task, which addressed sentiment analysis, abusive language detection, and hate speech detection in Tamil and Malayalam using multimodal data from YouTube. The best-performing team Rahman et al. (2024) combined ConvLSTM for video, BiLSTM for audio, and Naive Bayes for text for the abusive language detection subtask, demonstrating the effectiveness of ensemble methods. Imbwaga et al. (2024) proposed an audio-based approach for English and Kiswahili, highlighting the need for improved feature modeling in low-resource speech datasets. Transformer-based models have also been widely studied, with Sivanaiah et al. (2023) and M et al. (2023) evaluating var-

[1]Code for this work is available on GitHub

Language	Non-Hate (N)	Hate (H)				Total
		G (Gender)	P (Political)	R (Religious)	C (Personal Defamation)	
Tamil	287	63	33	61	65	509
Telugu	198	101	58	72	122	551
Malayalam	406	82	118	91	186	883

Table 1: Dataset description: The number of instances in each class (Non-Hate and Hate subcategories) for Tamil, Telugu, and Malayalam languages.

ious transformer and machine learning techniques for code-mixed Dravidian languages. Similarly, Sreelakshmi et al. (2024) analyzed multilingual embeddings for Dravidian languages, finding that MuRIL combined with SVM performed best while employing cost-sensitive learning to address class imbalance. These studies emphasize the importance of annotated datasets and specialized embeddings for effective hate speech detection in underrepresented languages.

Beyond text-based analysis, incorporating the speech modality has also proven to be crucial for hate speech detection, as vocal tone, prosody, and speech patterns provide additional context that aids in identifying offensive content. Bhesra et al. (2024) explored a multimodal framework combining MFCC audio features with text embeddings, employing a decision-level fusion strategy to improve detection accuracy. Similarly, Rana and Jha (2022) integrated audio-based emotion features with text-based semantic representations, demonstrating the benefits of cross-modal learning. Mandal et al. (2023) introduced a multimodal Transformer-based model, leveraging log mel spectrograms for speech and tokenized embeddings for text, with a novel "Attentive Fusion layer" to effectively combine both modalities. These works highlight the potential of multimodal architectures in detecting hate speech by leveraging complementary information from audio and text.

3 Dataset Description

In this study, we used the dataset provided by the organizers (Lal G et al., 2025; Anilkumar et al., 2024), comprising hate speech utterances collected from YouTube videos in three Indian languages: Malayalam, Tamil, and Telugu. The dataset includes both audio files (.wav format) and their corresponding text transcripts (.xlsx files). It is annotated with two primary classes: Hate (H) and Non-Hate (N). The Hate class is further categorized into the following subclasses based on the type of hate speech: Gender-based Hate (G), Political Hate (P), Religious Hate (R), and Personal Defamation (C). Upon verifying the dataset, we found that out of the 556 records in the Telugu .xlsx file, five corresponding audio files were missing, bringing the total number of matched audio-text pairs to 551. Similarly, in Tamil, out of 514 records, five audio files were missing, reducing the total available Tamil dataset to 509 multimodal instances. These final counts are reflected in Table 1, which reports the dataset distribution across languages.

4 Methodology

We implemented a multimodal dual Transformer-based framework for multiclass hate speech detection, drawing significant inspiration from Mandal et al. (2023), within the novel context of low-resource Indian languages and multiclass classification. An overview of the architecture is presented in Figure 1.

4.1 Data Preprocessing

Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from the audio data to capture frequency-related characteristics essential for detecting subtle variations in tone and speech patterns indicative of hate speech. Specifically, 40 MFCC features were computed on audio originally sampled at 44.1 kHz, with a Mel filterbank size (n_mels) of 128. Text data was tokenized to a maximum length of 128 tokens using the IndicBART tokenizer (Dabre et al., 2022).

4.2 Audio and Text Sampling

The audio and text data were then processed through dedicated sampling blocks. For audio data, the Speech Sampling Block involved passing the MFCC features through an LSTM layer to capture sequential dependencies, followed by the application of positional encodings to enhance temporal relationships. Specifically, the LSTM retains and updates hidden states over time via input, output,

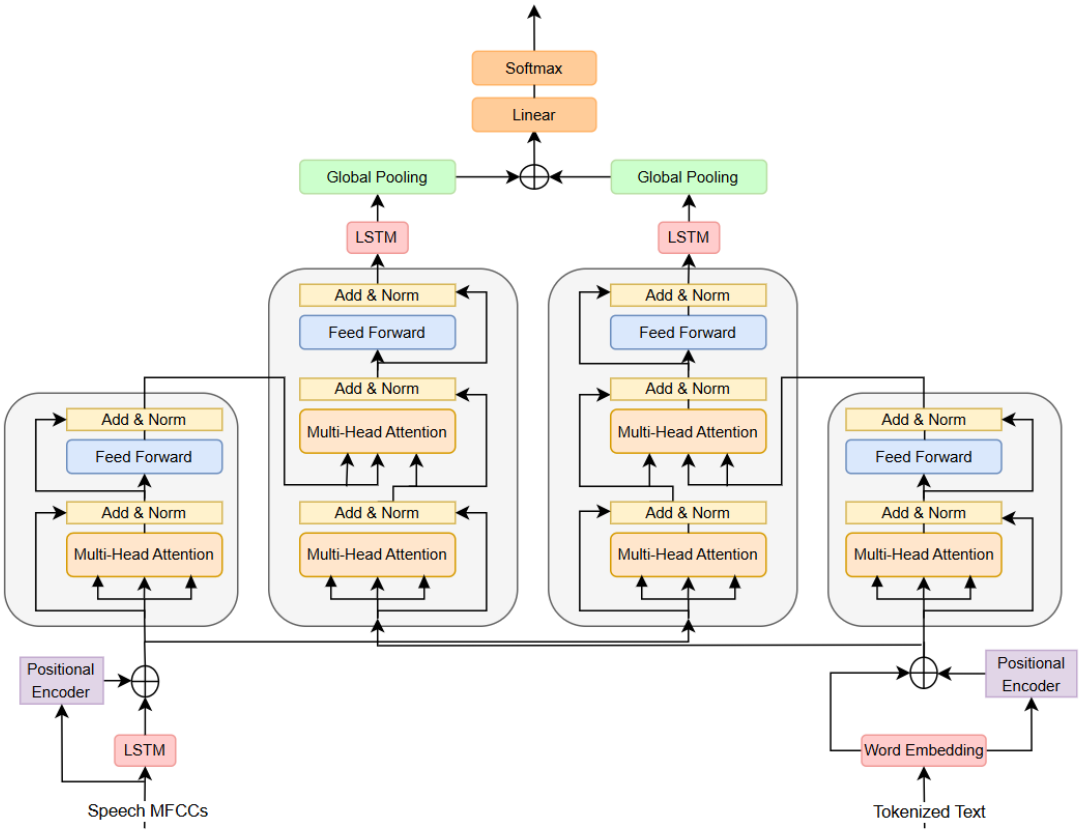


Figure 1: Model Architecture

and forget gates, allowing it to learn how consecutive MFCC frames connect. This gating mechanism leverages both short-term and long-term context, which enables effective modeling of continuous speech signals.

Similarly, the Text Sampling Block processed tokenized text sequences through an embedding layer and positional encodings. These sampling blocks ensured that both modalities were appropriately prepared for cross-modal interaction within the Transformer framework.

4.3 Dual Transformer Framework

The core architecture consists of two vanilla Transformers, introduced by Vaswani et al. (2017), each composed of two encoder and decoder layers with four attention heads (num_heads=4). The embedding size (d_model) was set to 128, and the feed-forward network (dff) had a dimension of 256. In the first transformer, the encoder processes speech features to extract contextual embeddings, which are then attended by the decoder, utilizing text features as input. This setup allows the decoder to learn from the contextual representations generated by the audio encoder. Conversely, in the second transformer, the encoder processes text features to produce contextual embeddings, while the decoder attends to these embeddings using audio features as input. This bidirectional mechanism facilitates effective cross-modal knowledge transfer, enabling each modality to enrich the other.

The outputs from both transformers are processed through separate LSTM layers, which capture long-term dependencies in the sequential data, enhancing the model's temporal understanding. These LSTM outputs then undergo global average pooling to reduce dimensionality while preserving essential information. The pooled features are subsequently concatenated, creating a unified representation of the audio and text modalities. Finally, this fused representation is passed through a dense layer with a softmax activation, classifying each input into one of five target categories.

4.4 Training Setup

The model was trained for 20 epochs using the Adam optimizer with a learning rate of 1e-4. To prevent overfitting, a dropout rate of 0.1 was applied, and an early stopping callback was employed with a patience value of 3.

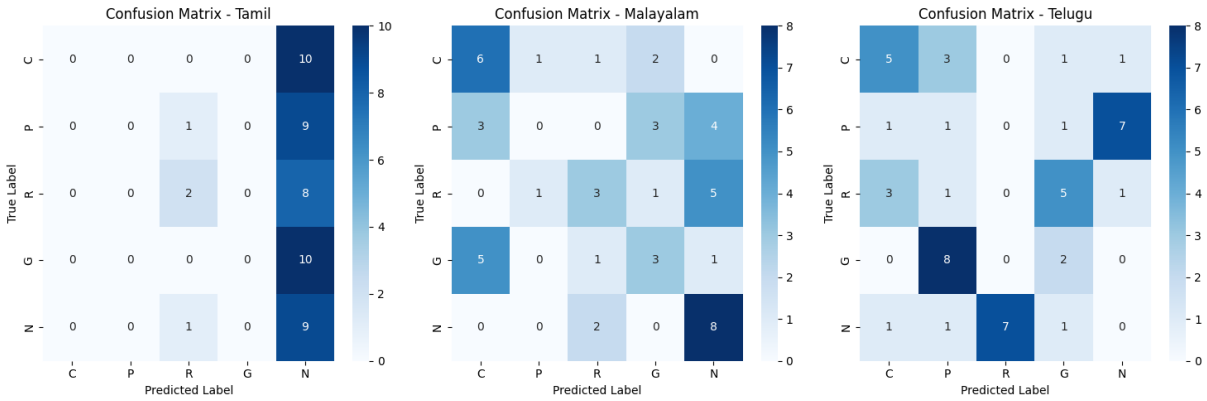


Figure 2: Confusion Matrices for Tamil, Malayalam, and Telugu Test Sets

5 Discussion of Results and Limitations

We submitted our runs for the Telugu and Malayalam subtasks, achieving macro-F1 scores of 0.1631 and 0.3480, respectively. These results placed us 14th in the Telugu subtask and 11th in the Malayalam subtask. Additionally, we conducted further experiments on the Tamil language. The macro-F1 scores for a subset of the dataset used for validation and the test set for each subtask are presented in Table 2. The variation in scores between the validation and test sets can be attributed to class imbalance. The training and validation sets contained a higher proportion of Non-Hate (N) instances, leading the model to overfit on the majority class. In contrast, the test sets had an equal distribution of 50 instances per class, exposing the model's difficulty in generalizing across underrepresented hate speech categories.

Subtask	Macro F1 Score (Val)	Macro F1 Score (Test)
Malayalam	0.6757	0.3480
Telugu	0.5146	0.1631
Tamil	0.2815	0.1271

Table 2: Macro F1 scores on validation and test sets for Malayalam, Telugu, and Tamil.

To further analyze model performance across languages, we present confusion matrices for Tamil, Malayalam, and Telugu test sets in Figure 2. These matrices provide insights into the distribution of misclassifications. The Tamil confusion matrix indicates a heavy bias toward predicting the 'N' (Non-Hate) class, resulting in a lack of diversity in predictions. Malayalam's confusion matrix, in contrast, shows a relatively balanced distribution with a few misclassifications spread across categories, indicating that the model is better at distinguishing between different classes but still struggles with borderline cases. For Telugu, the model misclassifies several instances of 'R' and 'N', suggesting that these categories are not well-separated in the learned representation.

While the model identified some patterns, its overall effectiveness was limited. The strong bias toward predicting the N (Non-Hate) class in Tamil and Telugu suggests difficulty in distinguishing hate speech categories due to insufficient minority class examples. In contrast, the slightly better performance in Malayalam highlights the impact of a larger dataset, though misclassifications persisted. These results emphasize the challenges of deep learning in low-resource languages, where data scarcity and class imbalance hinder accurate, nuanced hate speech detection.

6 Conclusion and Future Work

In this study, we explored a multimodal dual Transformer-based approach for hate speech detection in Dravidian languages. Our findings highlight the challenges of using deep learning for low-resource languages, where limited training data and class imbalance impact model performance. However, they also demonstrate the potential of multimodal approaches to leverage complementary audio and text information, offering promising directions for future advancements. We intend to explore alternative fusion and feature extraction techniques, class-weighting, and implement data augmentation strategies, such as backtranslation and Gaussian noise addition to enhance the model's ability to generalize and improve overall performance in future work.

References

Abhishek Anilkumar, Jyothish Lal G, B Premjith, and Bharathi Raja Chakravarthi. 2024. Dravlangguard: A multimodal approach for hate speech detection in dravidian social media. In *Speech and Language Technologies for Low-Resource Languages (SPELLL)*, Communications in Computer and Information Science.

Kirtilekha Bhesra, Shivam Ashok Shukla, and Akshay Agarwal. 2024. Audio vs. text: Identify a powerful modality for effective hate speech detection. In *The Second Tiny Papers Track at ICLR 2024*.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

Joan L Imbwaga, Nagatatna B Chittaragi, and Shashidhar G Koolagudi. 2024. Automatic hate speech detection in audio using machine learning algorithms. *International Journal of Speech Technology*, 27(2):447–469.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Hema M, Anza Prem, Rajalakshmi Sivanaiah, and Angel Deborah S. 2023. Athena@DravidianLangTech: Abusive comment detection in code-mixed languages using machine learning techniques. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 147–151, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Atanu Mandal, Gargi Roy, Amit Barman, Indranil Dutta, and Sudip Kumar Naskar. 2023. Attentive fusion: A transformer-based approach to multimodal hate speech detection. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 720–728, Goa University, Goa, India. NLP Association of India (NLPAI).

Arpan Nandi, Kamal Sarkar, Arjun Mallick, and Arkadeep De. 2024. A survey of hate speech detection in indian languages. *Social Network Analysis and Mining*, 14(1):70.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

Md. Rahman, Abu Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das, and Mohammed Moshiul Hoque. 2024. Binary_Beasts@DravidianLangTech-EACL 2024: Multimodal abusive language detection in Tamil based on integrated approach of machine learning and deep learning techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 212–217, St. Julian's, Malta. Association for Computational Linguistics.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *Preprint*, arXiv:2202.06218.

Rajalakshmi Sivanaiah, Rajasekar S, Srilakshmisai K, Angel Deborah S, and Mirnalinee ThankaNadar. 2023. Avalanche at DravidianLangTech: Abusive comment detection in code mixed data using machine learning techniques with under sampling. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 166–170, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Joseph B. Walther. 2022. Social media and online hate. *Current Opinion in Psychology*, 45:101298.

InnovateX@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Dravidian Languages

Moogambigai A, Pandiarajan D, Bharathi B
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
moogambigai2370071@ssn.edu.in
pandiarajan2370062@ssn.edu.in
bharathib@ssn.edu.in

Abstract

This paper presents our approach to the Shared Task on Detecting AI-Generated Product Reviews in Dravidian Languages as part of DravidianLangTech@NAACL 2025, as described by (Premjith et al., 2025). The task focuses on distinguishing between human-written and AI-generated reviews in Tamil and Malayalam, languages rich in linguistic complexities. Using the provided datasets, we implemented machine learning and deep learning models, including Logistic Regression (LR), Support Vector Machine (SVM), and BERT. Through preprocessing techniques like tokenization and TF-IDF vectorization, we achieved competitive results, with our SVM and BERT models demonstrating superior performance in Tamil and Malayalam respectively. Our findings underscore the unique challenges of working with Dravidian languages in this domain and highlight the importance of robust feature extraction.

1 Introduction

The proliferation of AI-generated content has brought both opportunities and challenges across various domains, including e-commerce, social media, and journalism. While AI can generate text efficiently, it also raises significant concerns regarding authenticity, particularly in online reviews (Kovács, 2024), where fake or AI-generated content can manipulate consumer trust and market dynamics. Detecting such content is essential for ensuring credibility and maintaining user trust in digital platforms (Diaz-Garcia and Carvalho, 2025).

This shared task focuses on detecting AI-generated product reviews in Dravidian languages, specifically Tamil and Malayalam (Priyadharshini et al., 2021). These languages pose unique challenges for natural language processing (NLP) due to their rich morphology, agglutinative nature, code-mixing tendencies, and lack of extensive annotated datasets. Tamil and Malayalam also frequently incorporate regional slang, making text analysis even more complex.

Our work addresses these challenges by employing both traditional machine learning methods, such as Support Vector Machine (SVM) and Logistic Regression (LR), and advanced deep learning approaches like BERT. We preprocess the data to capture linguistic nuances, leveraging techniques such as tokenization, stopword removal, and feature extraction using TF-IDF (Kumari et al., 2023). By comparing these models, we aim to identify systems that effectively distinguish human-written content from AI-generated text (Knight et al., 2023), while also contributing insights to the broader field of AI-generated content detection in low-resource languages.

Keywords

AI-generated content detection, Dravidian languages, Tamil and Malayalam, machine learning models, BERT, code-mixed text classification

2 Related Work

Detecting AI-generated content has been a growing area of research (Aho and Ullman, 1972), particularly in high-resource languages like English. Techniques such as transformer-based models (e.g., BERT) and traditional machine learning approaches (e.g., Support Vector Machines and Logistic Regression) have shown significant promise in identifying machine-generated text (Joshi et al., 2024). Studies utilizing BERT and its variants demonstrate strong performance in detecting patterns specific to AI-generated text (Shaik Vadla et al., 2024), leveraging contextual embeddings for improved classification accuracy. Traditional methods employing TF-IDF features combined with machine learning classifiers like SVM and Naive Bayes have also been effective in text classification

tasks, particularly in resource-constrained settings.

However, research on low-resource languages, such as Tamil and Malayalam, remains scarce despite their increasing presence in online spaces (American Psychological Association, 1983). Dravidian languages exhibit unique linguistic characteristics, including agglutination, rich morphology, and context-sensitive meaning, which make text processing challenging. Additionally, code-mixing and transliteration common in social media text add complexity to language modeling tasks (Abeera et al., 2023).

Prior work on Dravidian languages has primarily addressed sentiment analysis, sarcasm detection, and offensive language identification (Chandra et al., 1981). While these tasks share similarities with content classification, they do not specifically target the detection of AI-generated text (Ojo et al., 2024). Furthermore, the limited availability of annotated datasets and preprocessing tools tailored to Tamil and Malayalam constrains the applicability of standard NLP methods.

Building on these foundations (Abiola et al., 2025), this study investigates the applicability of both traditional machine learning models, such as Support Vector Machine (SVM) and Logistic Regression (LR), and deep learning approaches, including BERT and its multilingual variants, for detecting AI-generated product reviews in Tamil and Malayalam (Andrew and Gao, 2007). The research also considers the impact of linguistic characteristics such as code-mixing and transliteration (Rasooli and Tetreault, 2015), aiming to bridge the gap in AI-generated content detection for low-resource languages.

3 Dataset

The datasets provided in the shared task consisted of human-written and AI-generated product reviews in Tamil and Malayalam, structured with distinct features and labeled samples. The data distribution is presented in Table 1.

- **Training Data:** Tamil comprised 808 comments, while Malayalam contained 800, with features including ID, DATA, and LABELS.
- **Test Data:** Tamil included 100 comments, and Malayalam comprised 200, with features restricted to ID and DATA.

Comprehensive preprocessing was performed to standardize the datasets, including feature encoding, label normalization, and partitioning into training and evaluation subsets. This ensured optimal compatibility and performance across both traditional and transformer-based models (Javaji et al., 2024).

4 Methodology

Our approach to distinguishing human-written and AI-generated reviews in Tamil and Malayalam involved leveraging both traditional and transformer-based models. The methodology is detailed in the Figure 1:

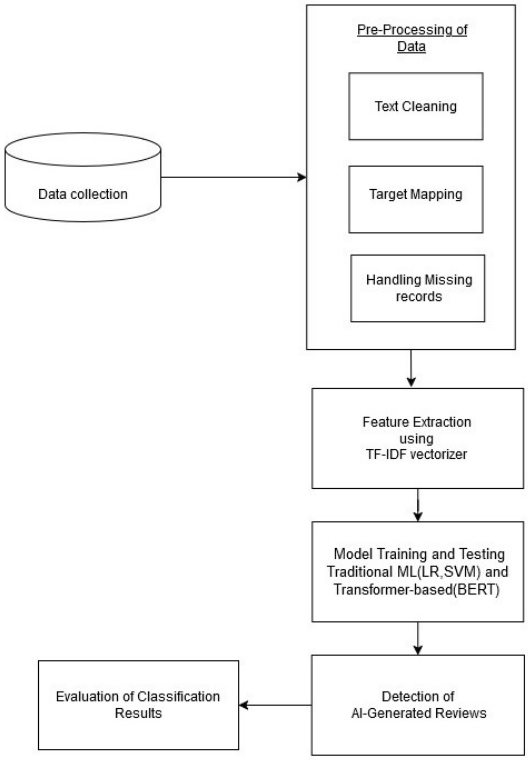


Figure 1: Framework of Proposed Methodology

4.1 Models for Classification

- **Logistic Regression (LR):** Utilized TF-IDF vectorization to transform textual data into numerical features, enabling efficient linear classification (Bhargav and Dhanalakshmi, 2024).
- **Support Vector Machine (SVM):** Combined robust preprocessing techniques with traditional classification to handle complex decision boundaries (Anbalagan et al., 2024).
- **BERT:** Leveraged the pre-trained transformer model with fine-tuning tailored separately for Tamil and Malayalam datasets, incorporating

Table 1: Data Distribution of Training and Testing Datasets

Language	Training Comments	Testing Comments
Tamil	808	100
Malayalam	800	200

advanced tokenization techniques (Ramachandruni et al., 2024).

4.2 Preprocessing Steps

To ensure data consistency and enhance model performance, the following preprocessing steps were undertaken:

- **Text Cleaning:** Removed special characters, normalized transliterated text, and addressed the challenges of code-mixing.
- **Feature Extraction:** Applied TF-IDF vectorization for LR and SVM to capture key textual patterns.
- **Tokenization:** Used BERT's subword tokenization to segment text into meaningful units for deep learning.

4.3 Training Process

The training process was designed to maximize the models' predictive capabilities:

- **Data Partitioning:** Split datasets into training and validation subsets, ensuring balanced representation of classes.
- **Model Optimization:** Conducted cross-validation and hyperparameter tuning to identify optimal configurations for each model.
- **Evaluation:** Monitored performance using accuracy and macro F1 metrics to ensure alignment with task objectives.

4.4 Deeper Analysis of SVM vs. BERT Performance in Tamil

Our experiments indicate that the SVM model outperforms BERT on Tamil data. Several linguistic and modeling factors contribute to this outcome:

- **Morphological Robustness:** Tamil's rich morphology benefits from SVM's TF-IDF n-gram representation, while BERT's subword tokenization may obscure semantics.
- **Code-Mixing and Transliteration:** SVM's bag-of-words approach is less affected by transliteration errors, whereas BERT struggles with out-of-vocabulary terms.
- **Dataset Limitations:** Tamil's limited dataset hinders BERT's fine-tuning, as its pretraining favors high-resource languages with standard orthography.

5 Experimental Results

The experimental evaluation reveals the performance of the classification models as summarized in Table 2. Among the models, SVM-Tamil achieved the highest accuracy (89.0%) and Macro F1 score (89.0%), demonstrating its robustness in classifying Tamil AI-generated and human-written reviews. LR-Tamil followed closely with an accuracy of 88.27% and an F1 score of 89.14%. in handling the intricate linguistic features of Tamil and Malayalam. BERT (Bala and Krishnamurthy, 2023) demonstrated competitive performance but faced challenges with code-mixed (Hande et al., 2021) and nuanced text.

The confusion matrix helps identify systematic misclassification trends.The confusion matrices corresponding to each model and language are presented in Figures 2 to 7, providing a detailed breakdown of predictions for human-written and AI-generated reviews.

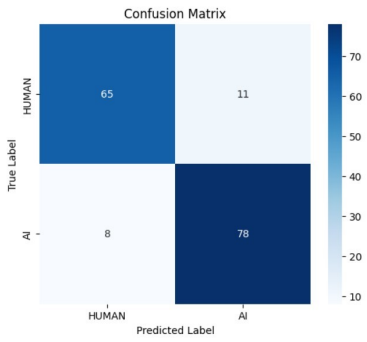


Figure 2: Confusion Matrix for LR Model (Tamil).

Table 2: Performance of models on the test dataset.

Model	Language	Accuracy	Precision	Recall	F1-Score
Logistic Regression (LR)	Tamil	88.27%	87.64%	90.70%	89.14%
Logistic Regression (LR)	Malayalam	76.88%	77.92%	75.00%	76.43%
Support Vector Machine (SVM)	Tamil	89.00%	89.00%	89.00%	89.00%
Support Vector Machine (SVM)	Malayalam	77.00%	77.00%	77.00%	77.00%
BERT	Tamil	78.00%	88.24%	62.50%	73.17%
BERT	Malayalam	79.01%	85.14%	73.26%	78.75%

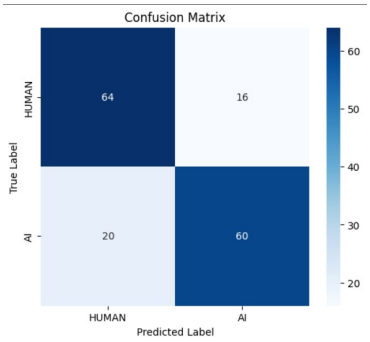


Figure 3: Confusion Matrix for LR Model (Malayalam).

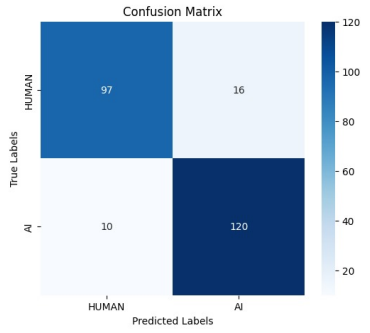


Figure 4: Confusion Matrix for SVM Model (Tamil).

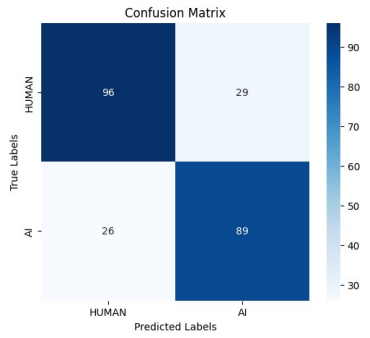


Figure 5: Confusion Matrix for SVM Model (Malayalam).

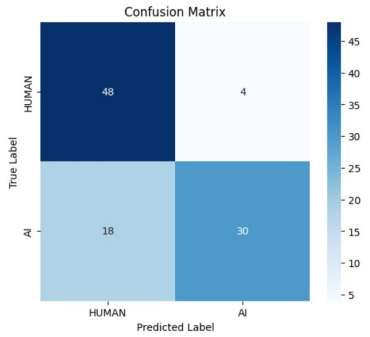


Figure 6: Confusion Matrix for BERT Model (Tamil).

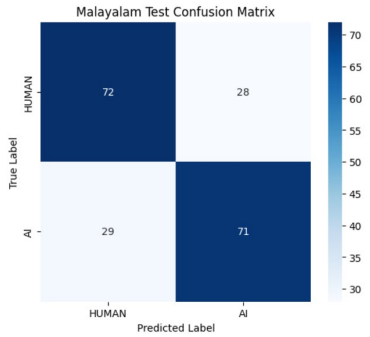


Figure 7: Confusion Matrix for BERT Model (Malayalam).

6 Conclusions

This study demonstrates the effectiveness of (Ando and Zhang, 2005) machine learning and (Bala and Krishnamurthy, 2023) deep learning models in detecting AI-generated content in Dravidian languages. BERT performed best for Malayalam, SVM for Tamil, and LR provided a strong baseline.

Future work will explore unsupervised and multilingual models to improve generalization in low-resource settings. This research advances AI-generated content detection in code-mixed languages.(Ignat et al., 2024)

For details, please visit the GitHub Repository.

7 Limitations

The model performed worse on Malayalam, achieving 79.01% accuracy with BERT, compared to Tamil, where the model reached 89.0% accuracy with SVM. Additionally, the model may misclas-

sify AI-generated text that closely mimics human writing. Another limitation is its difficulty in handling text that contains a mix of Tamil/Malayalam and English words, or text in Romanized script. Furthermore, with only approximately 800 samples per language, the model's generalization to unseen data is limited, particularly for new AI-generated or human-written reviews.

References

VP Abeera, Sachin Kumar, and KP Soman. 2023. Social media data analysis for malayalam youtube comments: Sentiment analysis and emotion detection using ml and dl models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51.

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide Ebenezer Ojo. 2025. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 262–270.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Akshatha Anbalagan, T Priyadharshini, A Niranjana, Shreedevi Balaji, and Durairaj Thenmozhi. 2024. Wordwizards@ dravidianlangtech 2024: Fake news detection in dravidian languages using cross-lingual sentence embeddings. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 162–166.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. Abhipaw@ dravidianlangtech: Fake news detection in dravidian languages using multilingual bert. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–238.

D Venkata Bhargav and R Dhanalakshmi. 2024. Performance analysis of logistic regression algorithm and random forest algorithm for predicting product review analysis. In *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pages 1–5. IEEE.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Jose A Diaz-Garcia and Joao Paulo Carvalho. 2025. A survey of textual cyber abuse detection using cutting-edge language models and large language models. *arXiv preprint arXiv:2501.05443*.

Adeep Hande, Karthik Puranik, Konthala Yasaswini, Ruba Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. 2021. Offensive language identification in low-resourced code-mixed dravidian languages using pseudo-labeling. *arXiv preprint arXiv:2108.12177*.

Oana Ignat, Xiaomeng Xu, and Rada Mihalcea. 2024. Maide-up: Multilingual deception detection of gpt-generated hotel reviews. *arXiv preprint arXiv:2404.12938*.

Prashanth Javaji, Pulaparthi Satya Sreeya, and Sudha Rajesh. 2024. Detection of ai generated text with bert model. In *2024 2nd World Conference on Communication & Computing (WCONF)*, pages 1–6. IEEE.

Ishika Joshi, Ishita Gupta, Adrita Dey, and Tapan Parikh. 2024. 'since lawyers are males..': Examining implicit gender bias in hindi language generation by llms. *arXiv preprint arXiv:2409.13484*.

Samsun Knight, Yakov Bart, and Minwen Yang. 2023. Generative ai and user-generated content: Evidence from online reviews. *Northeastern U. D'Amore-McKim School of Business Research Paper*, (4621982).

Balázs Kovács. 2024. The turing test of online reviews: Can we tell the difference between human-written and gpt-4-written online reviews? *Marketing Letters*, pages 1–16.

Kirti Kumari, Shirish Shekhar Jha, Zarikunte Kunal Dayanand, and Praneesh Sharma. 2023. Ml&ai_iiitranchi@ dravidianlangtech: Leveraging transfer learning for the discernment of fake news within the linguistic domain of dravidian language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 198–206.

Olumide E Ojo, Olaronke O Adebanji, Hiram Calvo, Alexander Gelbukh, Anna Feldman, and Ofir Ben Shoham. 2024. Doctor or ai? efficient neural network for response classification in health consultations. *IEEE Access*.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI

Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2021. Overview of the dravidiancodemix 2021 shared task on sentiment detection in tamil, malayalam, and kannada. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 4–6.

Indusree Ramachandruni, Sourav Mondal, Naga Venkata Mani Charan Jaladhi, Modukuri Sai Vyshnavi, Venkata Sai Sudheer Kumar Batchu, and Debnarayan Khatua. 2024. Enhancing product review authenticity detection with ensemble learning and bert model. In *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, pages 1–6. IEEE.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Mahammad Khalid Shaik Vadla, Mahima Agumbe Suresh, and Vimal K Viswanathan. 2024. Enhancing product design through ai-driven sentiment analysis of amazon reviews using bert. *Algorithms*, 17(2):59.

KSK@DravidianLangTech 2025: Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments Using Incremental Learning

Kalaivani K S[1], **Sanjay R**[1], **Thissyakkanna S M**[1], **Nirenjhanram S K**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
{kalaivani.cse, sanjayr.22aid}@kongu.edu
{thissyakkannasm.22aid, nirenjhanramsk.22aid}@kongu.edu

Abstract

The introduction of Jio in India has significantly increased the number of social media users, particularly on platforms like X (Twitter), Facebook, Instagram. While this growth is positive, it has also led to a rise in native language speakers, making social media analysis more complex. We took part in the shared task to classify political comments to classify social media comments from X (Twitter) into seven different categories. Tamil speaking users often communicate using a mix of Tamil and English, creating unique challenges for analysis and tracking. This surge in diverse language usage on social media highlights the need for robust sentiment analysis tools to ensure the platform remains accessible and user-friendly for everyone with different political opinions. In this study we trained four machine learning models, SGD Classifier, Random Forest Classifier, Decision Tree, and Multinomial Naive Bayes classifier to identify and classify the comments. Among these, the SGD Classifier achieved the best performance, with a training accuracy of 83.67% and a validation accuracy of 80.43%.

1 Introduction

In August 2024, India accounted for approximately 462 million active social media accounts, representing 32.2% of its population. This number is projected to grow exponentially, reaching nearly 1.2 billion users by 2029. With such massive growth, a significant portion of communication on these platforms is made by code-mixed language, an combination of native languages and English. Identifying politically inappropriate and sentimentally abusive comments within this data presents a unique and complex challenge due to the linguistic diversity and informal writing styles used by the users (Palit and Pal, 2018; Priyadharshini et al., 2021). Social media platforms possess immense power to influence public opinion in a short period, making it critical to monitor and classify political comments to ensure a safe space for users.(Kumar et al., 2021). Manual identification of such comments is not possible due to the volume of data and the complexity introduced by the surge of users and code-mixed language, which deviates substantially from standardized linguistic rules chakravarthi2020corpus. Advancements in machine learning (ML) and deep learning (DL) have enabled significant progress in NLP, particularly in analyzing multilingual and code-mixed datasets (Bojanowski et al., 2017a; Devlin et al., 2018; Chakravarthi, 2022). These technologies offer scalable, efficient solutions for text classification tasks. However, static models often struggle to adapt to the evolving nature of social media content, where trends, language usage, and sentiment expressions continuously change. To address this we have explored incremental learning to train the model. Incremental learning enables models to learn and adapt to new data without requiring complete retraining, making it highly effective for real-time applications like social media sentiment analysis (Parisi et al., 2019; Loshchilov and Hutter, 2017). In this study, we explore machine learning models for the multiclass sentiment analysis of Tamil X (formerly Twitter) comments. We categorize comments into seven sentiment classes: negative, neutral, none of the above, opinionated, positive, sarcastic, and substantiated. Our use incremental learning to efficiently adapt models to new data while maintaining performance on previously learned tasks. Various machine learning algorithms like Stochastic Gradient Descent (SGD) Classifier, Random Forest Classifier, Decision Tree, and Multinomial Naive Bayes Classifier are used. Out of 25 teams, our team ranked 18th place in the shared task(Chakravarthi et al., 2025).

2 Related Works

The study of sentiment analysis, particularly in the context of code-mixed text, has gained consider-

able attention in recent years due to the increasing prevalence of multilingual communication on social media platforms. Several researchers have explored various techniques to address the challenges of code-mixed data and sentiment classification.

(Khanuja et al., 2020) proposed a method for detecting offensive language in Hinglish (Hindi-English) code-mixed text using deep learning approaches. They highlighted the linguistic diversity and informal nature of code-mixed text as primary challenges, emphasizing the need for domain-specific embeddings.

ramesh2021dravidian focused on leveraging pre-trained multilingual language models like mBERT and XLM-R for sentiment analysis on Dravidian languages. Their experiments demonstrated the effectiveness of transfer learning for handling code-mixed languages while identifying areas for improvement in fine-tuning strategies

joshi2020lowresource explored sentiment classification in low-resource languages by combining rule-based and machine learning methods. They addressed the difficulties associated with data sparsity and introduced hybrid approaches that improved performance on small datasets

(Das et al., 2021) investigated the use of graph neural networks (GNNs) for sentiment analysis in code-mixed text. Their approach effectively captured relationships between words in multilingual sentences, offering a promising solution for sentiment classification tasks in linguistically diverse datasets.

(Ruder et al., 2019) provided a comprehensive survey on cross-lingual embeddings, discussing their applications for sentiment analysis and multilingual NLP tasks. They underscored the importance of shared embedding spaces for improving classification accuracy in code-mixed and low-resource scenarios.

Incremental learning techniques have also been explored in sentiment analysis to adapt to evolving data. (Bojanowski et al., 2017b) proposed continual learning algorithms designed to handle sequential tasks without catastrophic forgetting, making them highly suitable for real-time applications such as monitoring social media trends.Similarly,(Chen and Liu, 2018)introduced lifelong learning frameworks for text classification tasks, emphasizing the ability of models to accumulate knowledge across tasks.

These studies collectively highlight the advancements in machine learning and deep learning approaches for sentiment analysis on code-mixed text, while also emphasizing the potential of incremental learning to address dynamic and large-scale social media data. Building upon these works, this study adopts a combination of traditional classifiers and incremental learning techniques to classify Tamil X (formerly Twitter) comments into multiple sentiment categories.

3 Methodology

In this study machine learning is used to classify the training data into seven classes. This section discusses about various machine learning model used and procedures used in this study.

3.1 Dataset Used

The study uses the dataset provided by DravidianLangTech on Social Media sentiment classification (Chakravarthi et al., 2025). The dataset contains seven classses of different sentiment namely negative, neutral, none of the above, opinionated, positive, sarcastic, and substantiated. The training dataset contains 4533 rows of code-mixed tamil and the validation dataset contains 544 rows.

Label	Count
Opinionated	1,361
Sarcastic	790
Neutral	637
Positive	575
Substantiated	412
Negative	406
None of the above	171

Table 1: Training Data

Label	Count
Opinionated	153
Sarcastic	115
Neutral	84
Positive	69
Substantiated	52
Negative	51
None of the above	20

Table 2: Validation Data

3.2 Preprocessing Techniques

3.2.1 Removal of Hashtags, URLs, and Mentions:

Social media text often contains hashtags, URLs, and user mentions that do not offer any context and relevant meaning. The special character's such as '#' were removed and words such as 'www','.com' were removed keeping the content of hashtag and the URL's to train the model.

3.2.2 Whitespace Normalization:

Extra spaces which are present in the dataset were removed to ensure uniformity among the dataset.

3.2.3 Tokenization:

The cleaned up text is then split into words based on the occurence of space in between characters.

3.3 Models Used

In this study, we employed four machine learning algorithms to classify code-mixed Tamil text into seven sentiment categories: negative, neutral, none of the above, opinionated, positive, sarcastic, and substantiated. The models utilized are as follows:

3.3.1 Stochastic Gradient Descent (SGD) Classifier:

SGD is an optimization method that updates model parameters incrementally after evaluating each training example. This approach makes it efficient for large datasets and is particularly useful when a quick, approximate solution is acceptable.

3.3.2 Random Forest Classifier:

Random Forest is an ensemble learning method that builds multiple decision trees and combines their results to improve accuracy. Each tree in the forest makes a prediction, and the final output is determined by the majority vote among all trees. This method is effective for handling complex datasets with many features and is less prone to overfitting compared to individual decision trees.

3.3.3 Decision Tree Classifier:

A decision tree is a model that makes decisions by splitting data into subsets based on feature values, resembling a tree structure. It recursively splits the data at each node based on the feature that results in the best separation of classes. Decision trees are easy to understand and interpret, making them useful for problems where model transparency is important.

3.3.4 Multinomial Naive Bayes Classifier:

This probabilistic model is based on Bayes' theorem, assuming that features are conditionally independent given the class label. It calculates the probability of each class given the features and selects the class with the highest probability. Naive Bayes is particularly effective for text classification tasks, especially when the features (like words) are independent.

3.4 Training Methodology

The training data was divided into chunks of 108 batches, and the model was trained incrementally on these batches. This approach allows the model to learn from new data progressively, maintaining and building upon previous knowledge, which is particularly beneficial when dealing with large-scale datasets such as the one used in this study. Unlike traditional batch training, where the model is retrained from scratch with the entire dataset, incremental learning enables the model to update itself continuously without forgetting previously learned information. This is especially crucial in sentiment analysis, where language usage, expressions, and context evolve over time. By training in smaller chunks, the model adapts dynamically to emerging linguistic patterns and sentiment variations while preserving performance on earlier learned data. Additionally, this method reduces computational overhead, making real-time sentiment classification more efficient. Accuracy is monitored at each stage, ensuring that the model remains stable and effectively integrates new insights without suffering from catastrophic forgetting. Through continual updates, the model achieves improved generalization, making it well-suited for analyzing the ever-changing landscape of Tamil social media discourse.(dra, 2024)

4 Results and Discussion

Model performance can be assessed using a variety of metrics. In this study, we have chosen accuracy, precision, recall, and F1-score to evaluate the models. The machine learning models implemented include Stochastic Gradient Descent (SGD) Classifier, Random Forest Classifier, Decision Tree, and Multinomial Naive Bayes Classifier. The dataset is split into 100 chunks, and the models are trained in batches iteratively using incremental learning.

From Table 3, it is observed that the Stochastic Gradient Descent (SGD) Classifier achieves the

Model	Accuracy (%)	Validation Accuracy (%)
SGD Classifier	83.67	80.43
Naive Bayes	78.78	65.23
Logistic Regression	80.82	77.89
Random Forest Classifier	82.56	70.21

Table 3: Accuracy of different models on training and validation data

highest training accuracy of 83.67% and a validation accuracy of 80.43%, indicating strong generalization to unseen data. In contrast, the Naive Bayes model, with a validation accuracy of 65.23%, struggles to generalize, likely due to its assumption of feature independence, which does not hold in real-world linguistic data. The Logistic Regression model maintains balanced performance with 80.82% training accuracy and 77.89% validation accuracy, making it a stable alternative. The Random Forest Classifier, while achieving 82.56% training accuracy, exhibits a notable drop in validation accuracy (70.21%), suggesting overfitting due to its reliance on multiple decision trees.

While accuracy is an important evaluation metric, it does not fully capture the model's performance, particularly in imbalanced sentiment classes. Precision, recall, and F1-score provide deeper insights, revealing that Naive Bayes and Random Forest tend to misclassify minority sentiment categories, leading to lower recall scores. The use of incremental learning, where the dataset is processed in batches, enables the model to adapt to new linguistic trends in Tamil political discussions without catastrophic forgetting. This approach ensures the SGD Classifier maintains performance over time, making it well-suited for real-time sentiment analysis. Future improvements could explore weighted loss functions, data augmentation techniques, or deep learning-based models such as BERT or RoBERTa to enhance classification effectiveness in Tamil code-mixed sentiment analysis.

5 Conclusion

Sentiment analysis on code-mixed political data was conducted, and it was found that the Stochastic Gradient Descent (SGD) Classifier outperforms the other models by achieving a training accuracy of 83.67% and a validation accuracy of 80.43%. These results underscore the potential of SGD in handling code-mixed text in political sentiment analysis. In future work, further exploration can be carried out by incorporating more advanced techniques such as deep learning models, fine-tuning pre-trained models like BERT or RoBERTa for code-mixed data, and expanding the dataset to include more diverse political contexts to improve the model's robustness and performance. The code for our models and preprocessing methods is available here.

References

2024. *Proceedings of the EACL 2024 Workshop on Speech and Language Technologies for Dravidian Languages*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017a. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017b. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bharathi Raja Chakravarthi. 2022. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):75.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Zhiyuan Chen and Bing Liu. 2018. Lifelong learning for sentiment analysis tasks. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2204–2210.

S. Das et al. 2021. Graph neural networks for sentiment analysis in multilingual code-mixed text. *Knowledge-Based Systems*, 220:106901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Simran Khanuja, Kaustav Dey, El Moatez Billah Karim Nagoudi, et al. 2020. A new dataset and strong baselines for the detection of code-mixed offensive language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1719–1726.

Manish Kumar, Vikas Chauhan, Ashok Kumar Yadav, and Yogesh Kumar Meena. 2021. Multilingual sentiment analysis on social media: Challenges and applications. *Information Processing & Management*, 58(4):102509.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

N. Palit and K. Pal. 2018. Social media analytics: A survey on concepts, tools, and applications. *IEEE Access*, 6:12321–12345.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. Continual learning in deep neural networks: An empirical model. *Neural Networks*, 113:54–71.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Elizabeth Sherly, and John P. McCrae. 2021. Sentiment analysis in tamil-english code-mixed social media text. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(1):1–21.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual embeddings and their applications. *Journal of Artificial Intelligence Research*, 65:569–631.

BlueRay@DravidianLangTech-2025: Fake News Detection in Dravidian Languages

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2],
Aiswarya M[1], **Aruna T**[1], **Jeevaananth S**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{aiswaryam.22aid, arunat.22aid}@kongu.edu
{jeevaananths.22aid}@kongu.edu

Abstract

The rise of fake news presents significant issues, particularly for underrepresented languages. This study tackles fake news identification in Dravidian languages with two subtasks: binary classification of YouTube comments and multi-class classification of Malayalam news into five groups. Text preprocessing, vectorization, and transformer-based embeddings are all part of the methodology, including baseline comparisons utilizing classic machine learning, deep learning, and transfer learning models. In Task 1, our solution placed 17th, displaying acceptable binary classification performance. In Task 2, we finished eighth place by effectively identifying nuanced categories of Malayalam news, demonstrating the efficacy of transformer-based models.

1 Introduction

Fake news detection is a critical difficulty in combatting disinformation in today's digital landscape. Fake news is defined as intentionally misleading or incorrect material presented as legitimate news, which is typically designed to confuse readers and alter public opinion said by Anitha et al. (2024). In the view of Subramanian et al. (2025) proliferation of digital media has increased the dissemination of fake news, allowing misinformation to reach a large audience. Bala and Krishnamurthy (2023) highlight that fake news can take many forms, including manufactured tales, altered media, and biased content, especially on social media platforms where false narratives can quickly spread.

Devika et al. (2024) argue that misinformation adds to public panic, political polarization, and a reduction in faith in trustworthy news sources. Furthermore, unregulated fake news can sway public opinion, affect elections and policymaking, and incite social upheaval. Hariharan and Anand Kumar (2022) says detecting fake news is difficult due to the variety of writing styles, linguistic difficulties, and false news' ability to replicate actual information. Mohan et al. (2024) emphasize that standard detection methods frequently fail to capture contextual and cultural nuances, necessitating advanced natural language processing (NLP) models customized to these languages. According to Bade et al. (2024), machine learning and deep learning approaches are vital for developing robust false news detection models.

The shared task Fake News Detection in Dravidian Languages [1]aims on detecting fake news in underrepresented languages using binary and multi-class classification sets. This study describes a system for detecting fake news in various settings that uses text preprocessing, vectorization techniques (TF-IDF, BERT, etc.), advanced classification models such as transformers, and classic machine learning approaches. Section 2 summarizes works on detecting fake news, whereas Section 3 provides a full system description. Section 4 presents experimental results and analysis, followed by insights and future research directions.

2 Literature Review

Several studies have explored fake news detection in Dravidian languages, particularly Malayalam and Tamil, using various machine learning and deep learning approaches like Subramanian et al. (2023). Raja et al. (2023) proposed an optimized XLM-RoBERTa model, achieving improved accuracy in Malayalam fake news detection. Similarly, Sujan et al. (2023) introduced MalFake, a multimodal framework integrating Recurrent Neural Networks (RNNs) and VGG-16, demonstrating the effectiveness of combining text and images for misinformation identification. Coelho et al. (2023) adopted a traditional machine learning approach, experimenting with different classifiers for fake news detection. Eduri et al. (2023) ex-

[1]https://codalab.lisn.upsaclay.fr/competitions/20698

plored gradient accumulation-based transformer models, improving fake news classification performance in Malayalam. Additionally, Subramanian et al. (2024) provided an overview of the second shared task on fake news detection, highlighting key methodologies and benchmark datasets for Dravidian languages.

Other research efforts have focused on related NLP tasks for Malayalam and Tamil. Rameesa and Veeramanju conducted a systematic review on news headline categorization in Malayalam, addressing challenges in linguistic variations. Kumar et al. (2019) implemented deep learning-based part-of-speech tagging for Malayalam Twitter data, showcasing the importance of morphological analysis in NLP tasks. Ponnusamy et al. (2024) introduced an annotated dataset for misogyny detection in Tamil and Malayalam memes, emphasizing the role of social media in the spread of harmful narratives. Furthermore, YP and Nelliyullathil (2023) studied the spread of misinformation on Facebook, analyzing user engagement and the effectiveness of third-party fact-checking in curbing fake news. Farsi et al. (2024) improved MuRIL BERT, a multilingual BERT model designed for Indian languages, to classify fake news in Malayalam, with encouraging results. Rahman et al. (2024) used Malayalam-BERT, a language-specific transformer model, to classify fake news. They emphasized the relevance of domain-specific embeddings in increasing classification accuracy.

These studies collectively highlight the growing interest in fake news detection and NLP tasks in Dravidian languages Madhumitha et al. (2024). The advancements in transformers, multimodal learning, and traditional ML techniques have significantly contributed to improving detection accuracy, while challenges in code-mixing, linguistic diversity, and limited annotated datasets remain key areas for future research Osama et al. (2024).

3 Problem and System Description

The propagation of fake news on digital platforms has become a serious concern, fueling misinformation and upseting societal cohesion. This problem becomes more acute in multilingual populations, where code-mixed content hamper identification methods. Addressing this issue is critical to maintaining the credibility of the information shared online.

3.1 Dataset Description

The shared task dataset includes two subtasks with distinct structures:

Subtask 1 (Binary Classification): For this task the dataset has columns text and label. Column text refers to YouTube comments posted in Malayalam-English and label indicates if the comment is original or fake.

Subtask 2 (Multiclass Classification): For this task the dataset has columns Id, News, and Label. The Id is a unique identification given to each news story. Column News includes Malayalam news articles. Label sorts the news into five categories. The dataset is partitioned into two sets: training and testing.

Subtasks	Train	Test
Task 1	3,258	1020
Task 2	1901	200

Table 1: Dataset Description

3.2 Development Pipeline

Our system uses a systematic pipeline to detect fake news, which consists of the following stages: Text preprocessing, feature extraction, classification models, evaluation metrics. Figure 1 shows the workflow to detect fake news.

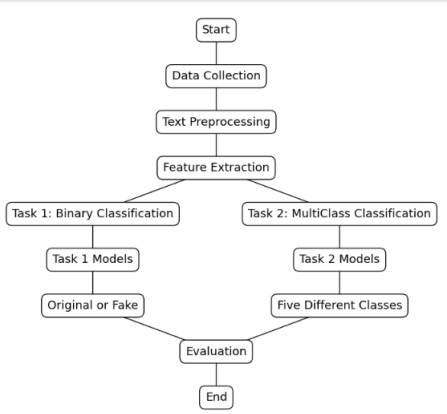


Figure 1: Proposed System Workflow.

3.2.1 Text Preprocessing

Text preparation is essential while creating Malayalam-English code-mixed YouTube comments and Malayalam news articles to detect fake news. To clean and organize the data efficiently, several techniques were required. When working with mixed-script tokens, the text was bro-

ken into words. Lowercasing and script normalisation ensured homogeneity. Stopwords and noise were removed using regular expression patterns, which included words, mentions, hashtags, emojis, and special characters. To preserve semantic meaning, words were stemmed and lemmatized in Malayalam to their base forms using language-specific procedures. Vectorization entailed transforming text into numerical representations using TF-IDF, Word2Vec, and transformer-based embeddings (BERT).

These preprocessing strategies ensured that models focused on meaningful content while decreasing noise and redundancy, resulting in higher classification accuracy.

3.2.2 Feature Extraction

Feature extraction translates text data into meaningful numerical representations, allowing for more successful fake news categorization. TF-IDF (Term Frequency-Inverse Document Frequency) emphasizes essential words while decreasing the influence of frequently used terms. Word Embeddings (Word2Vec) capture semantic links between words to improve contextual understanding, particularly in code-mixed text. Transformer-Based Embeddings (BERT) offers deep contextual meaning, improving classification accuracy for multilingual content.

These strategies aid the model's ability to discover patterns in both fake and authentic news, hence enhancing performance.

3.2.3 Classification Models

To efficiently recognize fake news in Dravidian languages, we used a variety of machine learning, deep learning models and transfer learning methods with various feature extraction methods designed to address the unique challenges of each task. Each model is briefly explained here, along with its performance.

Task 1: Binary Classification (Fake vs Original in Code-Mixed Youtube Comments)

SVM with TF-IDF: This model uses an optimal decision boundary to distinguish between fake and original news. Gradient Boosting Classifier with TF-IDF: This sequential learning approach corrects prior errors while detecting complicated patterns in false news. Logistic Regression with CountVectorizer: It trains the model using word frequency representation, resulting in successful text classification based on term occurrence patterns. Random Forest Classifier with Word2Vec: Uses word embeddings to capture semantic meaning, which improves classification accuracy.

Classification Model	Accuracy
SVM with TF-IDF	0.81
Gradient Boosting Classifier with TF-IDF	0.80
Logistic Regression with CountVectorizer	0.77
Random Forest Classifier with Word2Vec	0.65

Table 2: Accuracy of Binary Classification Models (Task 1).

The accuracy gained by these models is displayed in table 2 and the figure 2 shows the performance of SVM with TF-IDF model.

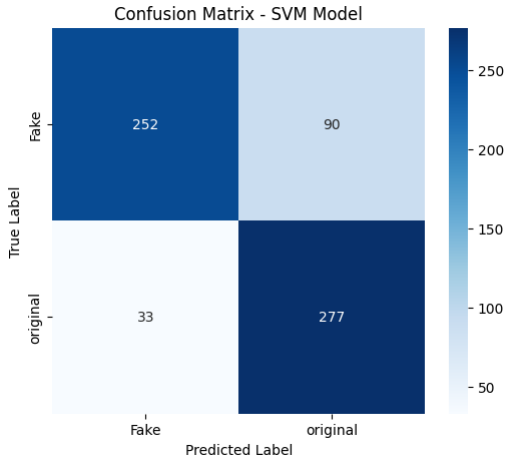


Figure 2: Performance of SVM with TF-IDF Model.

Task 2: Multiclass Classification (Classifying Malayalam News into Fake News Types)

Bi-LSTM: A deep learning model that extracts contextual meaning from both past and future words, improving classification accuracy for false news categories. XGBoost Classifier: It is a powerful boosting method that can handle imbalanced datasets and learn complex word associations. DistilBERT: Improves text comprehension through transformer-based contextual embeddings, resulting in high accuracy in fake news classification. SVM for Multiclass: Extends SVM for multiclass classification by specifying the boundaries between news categories.

The accuracy gained by these models is displayed in Table 3 and the figure 3 shows the performance of DistilBERT model.

Classification Model	Accuracy
DistilBERT	0.68
SVM for Multiclass	0.67
XGBoost Classifier	0.64
Bi-LSTM	0.54

Table 3: Accuracy of Multiclass Classification Models (Task 2).

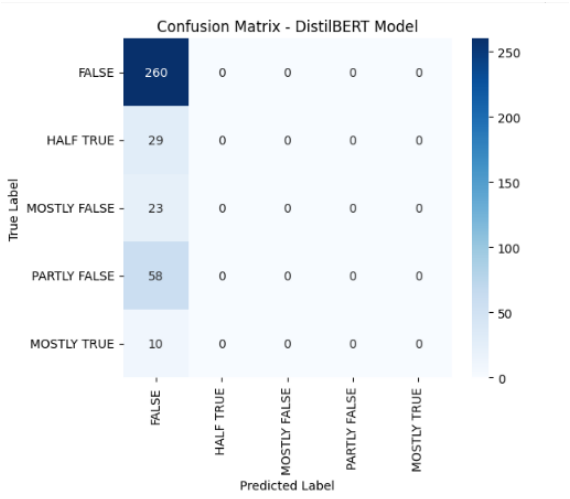


Figure 3: Performance of DistilBERT Model.

3.2.4 Evaluation Metrics

To ensure reliable fake news detection, the models are tested for accuracy, precision, recall, F1-score, macro F1-score, and loss. These measures helps determining the model's efficiency.

4 Experiments and Results

To classify fake news, experiments are conducted using various machine learning and deep learning models. For Task 1 (binary classification), SVM with TF-IDF attained a highest accuracy of 0.81 using a linear kernel, C value of 1.0, and balanced class weighting, demonstrating its effectiveness for Malayalam-English code-mixed comments. DistilBERT achieved 0.68 accuracy on Task 2 (multiclass classification) with a learning rate of 5e-5, batch size of 8, weight decay of 0.01, and three epochs, indicating its ability to classify nuanced Malayalam news. However, the confusion matrix revealed a bias toward the False category, necessitating class weighting and enhanced preprocessing to correct the class imbalance. Figure 4 and figure 5 shows their classification reports respectively.

Classification Report:

	precision	recall	f1-score	support
Fake	0.88	0.74	0.80	342
original	0.75	0.89	0.82	310
accuracy			0.81	652
macro avg	0.82	0.82	0.81	652
weighted avg	0.82	0.81	0.81	652

Figure 4: Classification Report of SVM with TF-IDF.

Classification Report:

	precision	recall	f1-score	support
FALSE	0.68	1.00	0.81	260
HALF TRUE	0.00	0.00	0.00	29
MOSTLY FALSE	0.00	0.00	0.00	23
PARTLY FALSE	0.00	0.00	0.00	58
MOSTLY TRUE	0.00	0.00	0.00	10
accuracy			0.68	380
macro avg	0.14	0.20	0.16	380
weighted avg	0.47	0.68	0.56	380

Figure 5: Classification Report of DistilBERT.

5 Conclusion

The purpose of this study was to detect fake news in Malayalam news articles and Malayalam-English code-mixed YouTube comments using various machine learning and deep learning algorithms. The study addressed issues such as code mixing, linguistic variances, and data scarcity while investigating successful categorization approaches. Our findings add to Dravidian language processing by comparing several ways to spotting disinformation. This Link contains the various algorithms used for this study. Future research can investigate data augmentation, multimodal techniques, and improved deep learning models to improve fake news identification.

6 Limitations

The results show that the model performs incorrectly when separating closely related classes in the multi-class classification problem, resulting in class overlap. Furthermore, while the binary classification performed well, it did occasionally misclassify borderline cases, demonstrating difficulties in dealing with subtle contextual distinctions.

References

R Anitha, S Navaneeth, Meharuniza Nazeem, and RR Rajeev. 2024. Code mixed english-malayalam sentiment analysis and sarcasm detection. In *2024*

15th International Conference on Computing Communication and Networking Technologies (ICCCNT), pages 1–6. IEEE.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024. Social media fake news classification using machine learning algorithm. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 24–29.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. Abhipaw@ dravidianlangtech: Fake news detection in dravidian languages using multilingual bert. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–238.

Sharal Coelho, Asha Hegde, G Kavya, and Hosahalli Lakshmaiah Shashirekha. 2023. Mucs@ dravidianlangtech2023: Malayalam fake news detection using machine learning approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 288–292.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Raja Eduri, Soni Badal, Borgohain Samir Kumar, and Lalrempuii Candy. 2023. Dravidian fake news detection with gradient accumulation based transformer model. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 466–471.

Salman Farsi, Asrarul Eusha, Ariful Islam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. Cuet_binary_hackers@ dravidianlangtech eacl2024: Fake news detection in malayalam language leveraging fine-tuned muril bert. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 173–179.

RamakrishnaIyer LekshmiAmmal Hariharan and Madasamy Anand Kumar. 2022. Impact of transformers on multilingual fake news detection for tamil and malayalam. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 196–208. Springer.

S Kumar, M Anand Kumar, and KP Soman. 2019. Deep learning based part-of-speech tagging for malayalam twitter data (special issue: deep learning techniques for natural language processing). *Journal of Intelligent Systems*, 28(3):423–435.

M Madhumitha, M Kunguma, J Tejashri, et al. 2024. Techwhiz@ dravidianlangtech 2024: Fake news detection using deep learning models. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 200–204.

Aiswarya Mohan, Nafla Iqbal, and Manaal Mashpher. 2024. Malayalam fake news detection using optimized convolutional neural network (opcnn). In *2024 11th International Conference on Advances in Computing and Communications (ICACC)*, pages 1–6. IEEE.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Cuet_nlp_goodfellows@ dravidianlangtech eacl2024: A transformer-based approach for detecting fake news in dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192.

Rahul Ponnusamy, Kathiravan Pannerselvam, R Saranya, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, S Bhuvaneswari, Anshid Ka, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in tamil and malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488.

Tanzim Rahman, Abu Raihan, Md Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. Cuet_duo@ dravidianlangtech eacl2024: Fake news classification using malayalam-bert. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 223–228.

Eduri Raja, Badal Soni, and Sami Kumar Borgohain. 2023. nlpt malayalm@ dravidianlangtech: Fake news detection in malayalam using optimized xlm-roberta model. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 186–191.

K Rameesa and KT Veeramanju. A systematic review on various approaches for news headlines categorization in malayalam language.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task

on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Adhish S Sujan, Aleena Benny, VS Anoop, et al. 2023. Malfake: A multimodal fake news identification for malayalam using recurrent neural networks and vgg-16. *arXiv preprint arXiv:2310.18263*.

Habeeb Rahman YP and Muhammadali Nelliyullathil. 2023. Spread of misinformation in malayalam: A case study on the user engagement and impact of third-party fact-checking on facebook.

KEC_AI_ZEROWATTS@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2],
Naveenram CE[1], **Vishal RS**[1], **Srinesh S**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{naveensrn1935,rsvishaltpr,srisiva262005}@gmail.com

Abstract

Hate speech detection in code-mixed Dravidian languages presents significant challenges due to the multilingual and unstructured nature of the data. In this work, we participated in the shared task to detect hate speech in Tamil, Malayalam, and Telugu using both text and audio data. We explored various machine learning models, including Logistic Regression, Ridge Classifier, RandomForest, and Convolutional Neural Networks (CNN). For Tamil text data, Logistic Regression achieved the highest macro-F1 score of 0.97, while Ridge Classifier performed best for audio with 0.75. In Malayalam, Random Forest excelled for text with 0.97, and CNN for audio with 0.69. For Telugu, Ridge Classifier achieved 0.89 for text and CNN 0.87 for audio.These results demonstrate the efficacy of our multimodal approach in addressing the complexity of hate speech detection across the Dravidian languages.Tamil:11th rank,Malayalam :6th rank,Telugu:8th rank among 145 teams

1 Introduction

Hate speech on social media is becoming more and more troublesome, particularly in multilingual situations where users mix imported terms with native scripts, such as Telugu, Tamil, and Malayalam. Because there are few annotated datasets and linguistic variation, it is difficult to detect such speech. To improve the accuracy of detection, this study suggests a multimodal strategy that combines text and audio features. Random Forest, Ridge Classifier, and Logistic Regression are examples of text-based models that are used to assess linguistic clues. Convolutional Neural Networks (CNNs) are used to process audio inputs in order to extract prosodic.

The shortcomings of conventional text-only approaches are addressed by combining text and audio predictions. The study shows that using both modalities greatly enhances detection performance by using YouTube data. The potential of multimodal systems is demonstrated by CNNs' efficacy in audio analysis and text machine learning models. This method provides a strong framework for spotting hate speech in intricate, multilingual internet settings.

2 Literature Survey

Barman and Das (2023) developed multimodal models for abusive language detection and sentiment analysis in Tamil and Malayalam. They used MFCC for audio, ViT for images, and mBERT for text, achieving a weighted F1 score of 0.5786 in abusive language detection and securing first place.

Bala and Krishnamurthy (2023) addressed sentiment analysis in Tamil and Malayalam videos and the detection of abusive language in Tamil multimodal videos. Their models used MViT for video, OpenL3 for audio, and BERT for text, demonstrating effective multimodal fusion.

Rahman et al. (2024) applied a multimodal strategy integrating text, audio, and video for Tamil abusive language detection. They used ConvLSTM, 3D-CNN, and a hybrid 3D-CNN+BiLSTM for video, and combined textual predictions from MNB, LR, and LSTM with audio features using a late fusion model. Their best model (ConvLSTM+BiLSTM+MNB) achieved a macro F1 score of 71.43, ranking first in the task.

Premjith et al. (2024) summarized the results of a shared task on multimodal sentiment analysis, abusive language detection, and hate speech detection in Tamil and Malayalam. Despite 39 teams participating, only two submitted results, which were evaluated using macro F1-score.

Anierudh et al. (2024) focused on three tasks: (1) sentiment classification in Tamil and Malayalam (highly positive, positive, neutral, negative, highly negative); (2) abusive language detection in Tamil; (3) hate speech detection in Tamil (Caste, Offen-

sive, Racist). They used machine learning models and oversampling strategies to handle dataset biases.

Rajalakshmi et al. (2024) addressed hate speech detection in code-mixed languages using transliteration. They achieved F1 scores of 0.68 (Logistic Regression) and 0.70 (Bi-GRU), contributing to research in preventing hate speech in mixed-language content.

Sreelakshmi et al. (2024) explored hate speech and offensive language (HOS) detection in Tamil-English, Malayalam-English, and Kannada-English using multilingual transformer-based embeddings. MuRIL performed best across datasets, achieving 96 accuracy in Malayalam and 72 in Tamil (DravidianLangTech 2021), and 76 in Tamil and 68 in Malayalam (HASOC 2021). A new annotated Malayalam-English test set was also introduced.

Yasaswini et al. (2021) worked on offensive language detection in Malayalam, Tamil, and Kannada at EACL 2021. They categorized social media posts into six classes using transfer learning. Their source code was released publicly.

3 Task Description

This study focuses on multimodal hate speech detection in Tamil, Malayalam, and Telugu using a dataset sourced from YouTube videos.The dataset consists of audio and text samples that have been categorized as either non-hate or hate (subclasses: political, religious, gender, and personal defamation). Vectorizers such as Count Vectorizer, TF-IDF, and Word2Vec were used to handle text data, while pre-processing was done on audio data to extract prosodic characteristics. Text was subjected to machine learning models such Random Forest, Ridge Classifier, and Logistic Regression, while audio was subjected to CNN. The advantages of a multimodal strategy were demonstrated by evaluating these models' performance using the macro-F1 score.Lal G et al. (2025) Tamil secured the 11th rank, Malayalam secured the 6th rank, and Telugu secured the 8th rank among 145 teams.

4 Dataset Description

4.1 Text Data Description

The Text dataset consists of three languages: Malayalam, Tamil, and Telugu, with each record labeled as either Non-Hate or Hate. Content that does not include offensive language is categorized as Non-Hate (abbreviated "N"), whereas content that falls within the Gender (G), Political (P), Religious (R), and Personal Defamation (C) categories is grouped together into the Hate category. Smaller test sets are available, however the training dataset consists of 883 records in Malayalam, 1397 records in Tamil, and 1953 records in Telugu. The training data for each language is broken out in depth in Table 1 below, which displays the distribution of the Non-Hate and Hate categories. With an emphasis on hate speech detection across many languages, this dataset is intended to train algorithms that categorize material as either harmful or non-harmful.

Language	Non-Hate(N)	Hate(C,G,P,R)
Malayalam	406	477
Tamil	287	491
Telugu	198	175

Table 1: Dataset Description of Text-Train

4.2 Audio Data Desciption

The Audio dataset is organized similarly to the Text dataset, with entries classified as either Hate or Non-Hate. While hateful content falls under the categories of gender (G), politics (P), religion (R), and personal defamation (C), non-hateful content is audio that does not contain damaging speech. For training, the Audio-Train dataset consists of 883 Malayalam, 509 Tamil, and 551 Telugu recordings, along with smaller test sets. The training data for each language is broken out in depth in Table 2 below, which displays the distribution of the Non-Hate and Hate categories. With an emphasis on identifying hate speech in many languages, this dataset is used to train algorithms for identifying damaging speech in audio data.

Language	Non-Hate(N)	Hate(C,G,P,R)
Malayalam	406	477
Tamil	287	222
Telugu	198	353

Table 2: Dataset Description of Audio-Train

5 Methodology

5.1 System Architecture

There are two pipelines in the system: audio and text. TF-IDF and Count Vectorizer are used for preprocessing, tokenization, and vectorization in

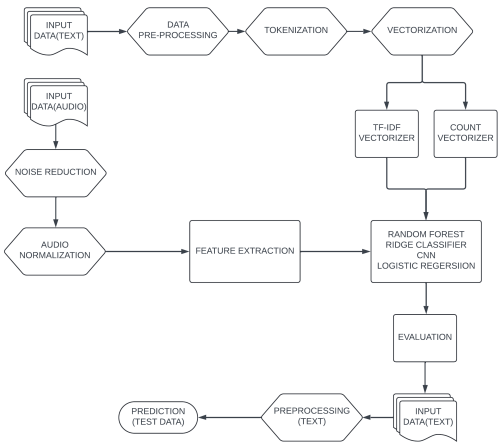


Figure 1: System Architecture

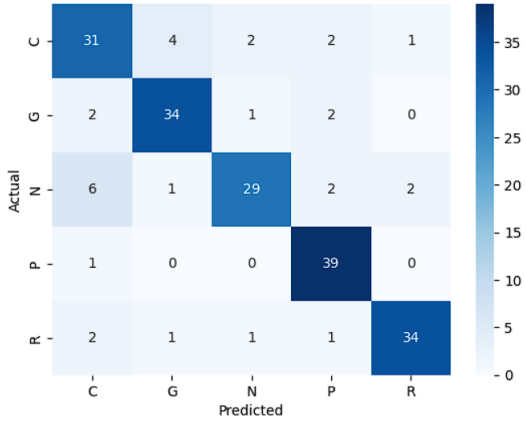


Figure 2: Confusion Matrix of Tamil-Text

the text pipeline. Random Forest, Ridge Classifier, and Logistic Regression are then used for classification. CNN-based classification, noise reduction, and normalization are all part of the audio pipeline, and the results are combined to produce the final forecast..

5.2 Data Preprocessing

In order to create numerical representations suitable for machine learning models, the dataset underwent specific preprocessing steps tailored for each modality: for text data, we removed unnecessary punctuation, URLs, and symbols; for audio data, we segmented the audio into smaller chunks, normalized the amplitude, and reduced background noise; for audio data, we extracted prosodic features like pitch, energy, and spectral characteristics to capture tonal and temporal information relevant to hate speech detection; and for speech data, we tokenized the text into individual words, followed by stop word removal and stemming/lemmatization to reduce words to their base forms.

5.3 Model Development

We used a range of deep learning and machine learning algorithms to classify hate speech. We employed Ridge Classifier, Random Forest, and Logistic Regression for text data because of their efficacy and interpretability with high-dimensional text data. We applied Convolutional Neural Networks (CNN) to audio data in order to identify tone and temporal patterns. To take linguistic and cultural quirks into consideration, each model was taught independently for Telugu, Tamil, and Malayalam. Data imbalance was addressed by class balancing approaches, and model performance was optimized by hyperparameter tweaking.

6 Performance Evaluation

The performance of the models was evaluated using the Macro-F1 score. Table 3 summarizes the results for text and audio modalities across Tamil, Malayalam, and Telugu. GitHub Repository: Dravidan-LangTech

Language	Modality	Macro-F1 Score
Tamil	Text	0.97
Tamil	Audio	0.75
Malayalam	Text	0.97
Malayalam	Audio	0.69
Telugu	Text	0.89
Telugu	Audio	0.87

Table 3: Performance Metrics for Text and Audio Modalities

6.1 Tamil

Logistic Regression proved highly effective in categorizing hate speech in Tamil text, achieving a Macro-F1 score of 0.97. Using Count Vectorizer, TF-IDF, and Word2Vec techniques, the model accurately distinguished between Hate and Non-Hate categories, effectively identifying themes like gender, politics, religion, and personal defamation based on language patterns.

For audio data, the Ridge Classifier outperformed other models with a Macro-F1 score of 0.75, highlighting the importance of speech spectral features in detecting hate speech. While CNN was successful in capturing speech's temporal aspects, it performed less effectively compared to other classifiers for audio-based detection.

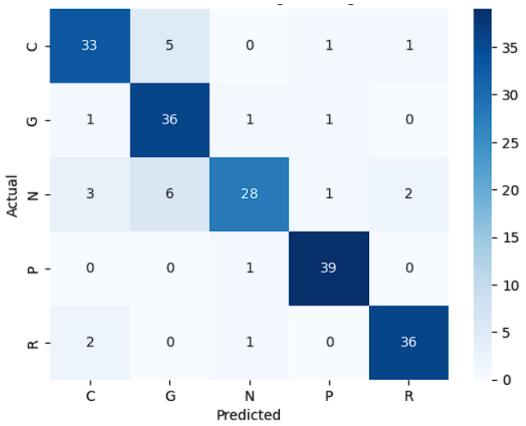


Figure 3: Confusion Matrix of Malayalam-Text

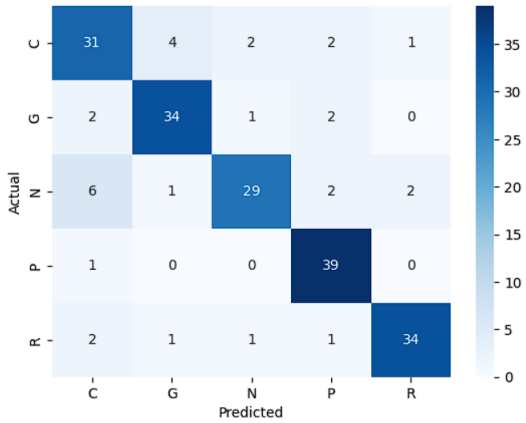


Figure 4: Confusion Matrix of Telugu-Text

6.2 Malayalam

The Random Forest Classifier had the greatest Macro-F1 score of 0.97 for the text modality in Malayalam. This outcome shows how well the model can categorize hate speech from a variety of subclasses, including gender, political, religious, and personal defamation. By utilizing the various variables that were recovered using vectorization approaches, the Random Forest model—an ensemble approach—provided excellent prediction performance.For Malayalam data, the confusion matrix for the top-performing model is shown in Figure 3.

CNN's Macro-F1 score for the audio modality was 0.69, which is less than the text performance but still shows a respectable level of success in identifying tonal characteristics linked to hate speech in Malayalam. The difficulties in utilizing CNN to analyze the rich prosodic elements of Malayalam speech may be the cause of the worse results.

6.3 Telugu

For Telugu, the text modality, the Ridge Classifier received a Macro-F1 score of 0.89. This shows that the model successfully distinguished between the Non-Hate and Hate categories, including their several subclasses, and was quite successful in detecting hate speech in Telugu. For the model to function well, the vectorized features from Word2Vec and TF-IDF were essential.The confusion matrix for the top-performing model using Telugu data is shown in Figure 4.

For the audio modality, With a Macro-F1 score of 0.87, CNN fared better than other models, demonstrating the model's capacity to accurately represent Telugu speech dynamics. CNN's excellent performance in audio categorization demonstrates its capacity to examine speech patterns and detect hostile or aggressive behavior.

7 Limitations

Although our multimodal method for identifying hate speech in Dravidian languages yields encouraging findings, there are a number of drawbacks to take into account. First, the dataset size is minimal, which could restrict how broadly the models can be applied. Second, prosodic features could be more effectively captured by further optimizing the audio data preparation pipeline. The models can also have trouble handling code-mixed content, which is prevalent on social media. Enhancing the integration of text and audio modalities and growing the dataset should be the main goals of future research.

8 Conclusion

This study successfully combined text and audio data using multimodal approaches to identify hate speech in Telugu, Tamil, and Malayalam. Text-based classification yielded strong Macro-F1 scores for machine learning models such as Random Forest and Logistic Regression. With a score of 0.87 in Telugu, CNN performed exceptionally well in audio, while Malayalam fared marginally worse. The findings emphasize how crucial prosodic characteristics are for identifying hate speech. Combining deep learning with more conventional machine learning techniques showed promise. To increase the accuracy of multilingual hate speech detection, future developments might concentrate on feature extraction optimization and model calibration.

References

S Anierudh, Abhishek R, Ashwin Sundar, Amrit Krishnan, and Bharathi B. 2024. Wit hub@DravidianLangTech-2024:multimodal social media data analysis in Dravidian languages using machine learning models. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 229–233, St. Julian's, Malta. Association for Computational Linguistics.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. AbhiPaw@DravidianLangTech: Multimodal abusive language detection and sentiment analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 140–146, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Shubhankar Barman and Mithun Das. 2023. hate-alert@ dravidianlangtech: Multimodal abusive language detection and sentiment analysis in dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 217–224.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, Bharathi B, Saranya Rajiakodi, Rahul Ponnusamy, Jayanth Mohan, and Mekapati Reddy. 2024. Findings of the shared task on multimodal social media data analysis in Dravidian languages (MSMDA-DL)@DravidianLangTech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61, St. Julian's, Malta. Association for Computational Linguistics.

Md. Rahman, Abu Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das, and Mohammed Moshiul Hoque. 2024. Binary_Beasts@DravidianLangTech-EACL 2024: Multimodal abusive language detection in Tamil based on integrated approach of machine learning and deep learning techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 212–217, St. Julian's, Malta. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Saptharishee M, Hareesh S, Gabriel R, and Varsini Sr. 2024. DLRG-DravidianLangTech@EACL2024 : Combating hate speech in Telugu code-mixed text on social media. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 140–145, St. Julian's, Malta. Association for Computational Linguistics.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194, Kyiv. Association for Computational Linguistics.

MNLP@DravidianLangTech 2025: A Deep Multimodal Neural Network for Hate Speech Detection in Dravidian Languages

Shraddha Chauhan
Department of ECE
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
shraddha.20224147@mnnit.ac.in

Abhinav Kumar
Department of CSE
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
abhik@mnnit.ac.in

Abstract

Social media hate speech is a significant issue because it may incite violence, discrimination, and social unrest. Anonymity and reach of such platforms enable the rapid spread of harmful content, targeting individuals or communities based on race, gender, religion, or other attributes. The detection of hate speech is very important for the creation of safe online environments, protection of marginalized groups, and compliance with legal and ethical standards. This paper aims to analyze complex social media content using a combination of textual and audio features. The experimental results establish the effectiveness of the proposed approach, with F_1-scores reaching 72% for Tamil, 77% for Malayalam, and 36% for Telugu. Such results strongly indicate that multimodal methodologies have significant room for improvement in hate speech detection in resource-constrained languages and underscore the need to continue further research into this critical area.

1 Introduction

Social media has changed the way people communicate, share information, and express opinions (Saumya et al., 2024; Bhawal et al., 2021). This digital transformation has been of great benefit, but it has also enabled the spread of hate speech, misinformation, and harmful content (Saumya et al., 2024; Biradar et al., 2022). Hate speech on social media is a serious issue since it encourages discrimination, hostility, and violence, thereby negatively affecting individuals and communities (Saumya et al., 2021; Kumar et al., 2021). Such content identification and blocking are highly pertinent to ensuring safer online surroundings within multilingual regions with strong multicultural influences.

Dravidian languages, such as Tamil, Malayalam, and Telugu, represent a rich linguistic heritage spoken by millions in South India and neighboring countries. Despite their prominence, these languages remain underrepresented in computational linguistics and natural language processing (NLP) research. The unique linguistic features of Dravidian languages, such as their morphology, syntax, and phonology, make it difficult to analyze texts automatically. Moreover, hate speech in these languages is often code-mixed, using native words with English, and auditory cues in multimedia formats like speech.

This study develops hate speech detection systems for Dravidian languages by adopting a multimodal approach that incorporates both textual and auditory data. Unlike traditional text-only methods, the integration of audio features enables the detection of tonal, emotional, and contextual cues that are critical for identifying hate speech in spoken communication. By leveraging advanced deep learning techniques and multimodal data fusion, our work aims to improve the accuracy and reliability of hate speech detection in Tamil, Malayalam, and Telugu social media posts.

The rest of the structure of the paper follows the sequel: Section 2 lists the related work, Section 3 deals with the dataset and task, and Section 4 details the proposed methodology. Section 5 reports results from the proposed model, Section 6 concludes the paper, and Section 7 details the limitations of proposed model.

2 Related Work

Hate speech refers to offensive or prejudiced communication aimed at demeaning individuals or groups based on their identity or beliefs (Kumar et al., 2020; Mishra et al., 2020). (Sharma et al., 2024) provided an extensive survey relating to the study done on hate speech detection in South Asian languages and classified all studies based on their tasks, datasets, and methodologies. (Diaz-Garcia and Carvalho, 2025) provided an in-depth survey on the impact of new technologies, such as Language Models and Large Language Models, on the

evolution of abusive content detection.

(Sreelakshmi et al., 2024a) introduced a cost-sensitive learning approach towards hate speech and offensive language detection in code-mixed text. This improves methodologies for solving complex linguistic problems of Dravidian languages. (Premjith et al., 2024) significantly contributed to the research area of hate speech detection in Dravidian languages by performing two major tasks. They highlighted cutting-edge strategies for handling hate speech on social media platforms by presenting the results of the "HOLD-Telugu" shared task on the identification of hate and offensive language in Telugu code-mixed text. They also participated in the "MSMDA-DL" shared task, which focused on multimodal social media data analysis by combining audio and textual elements to enhance the identification of hate speech in Dravidian languages.

(Roy and Kumar, 2025) proposed a cost-sensitive learning approach for detecting code-mixed hate speech in social media posts, leveraging advanced multilingual models and machine learning classifiers to address the challenges of linguistic diversity in Dravidian languages. (Yuan and Rizoiu, 2025) proposed a multi-task learning approach to improve the generalization of hate speech detection models on different datasets. In this respect, they introduces the PubFigs dataset to analyze hate speech in political discourse.

(Singh et al., 2025) proposed a multimodal approach incorporating emotional understanding to detect offensive content, focusing on women's harassment. (Raphel et al., 2024) explored the use of multilingual transformer-based embedding models and machine learning classifiers to detect hate speech and offensive language in code-mixed Dravidian language texts.

(Sai et al., 2024) focused on breaking linguistic barriers by developing robust methodologies to detect hate speech in Telugu-English code-mixed text. Studies such as (Kakati and Dandotiya, 2024) explored ensemble methods for improved accuracy. Using multilingual BERT models, works such as (Zamir et al., 2024) and (Abitte Kanta et al., 2024) concentrated on detecting hate speech and offensive language in Telugu.

In the Malayalam code-mixed language, (Sreelakshmi et al., 2024b) presented a deep learning-based feature fusion technique designed for sentiment analysis and offensive text identification. Through the integration of several linguistic characteristics, the study emphasizes the difficulties presented by code-mixed texts and the effectiveness of sophisticated deep learning models in precisely identifying objectionable material and sentiment.

Traditional machine learning models, such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995), K-Nearest Neighbors (KNN) (Cover and Hart, 1967), Random Forest (RF) (Breiman, 2001), Naive Bayes (NB) (John and Langley, 1995), and Logistic Regression (LR) (Hosmer and Lemeshow, 1989), were widely used for classification tasks, including hate speech detection (Boishakhi et al., 2021).

3 Dataset & Task

The Tamil, Malayalam, and Telugu datasets consist of 509, 883, and 551 audio and the corresponding transcript for training and testing data consists of 50 audio and the corresponding transcript for each of the languages (Lal G et al., 2025). The audio and transcripts are classified into Hate and Non-Hate categories. The hate class is further divided into subclasses: Gender (G), Political (P), Religious (R), and Personal Defamation (C), while the Non-Hate class is denoted as N.

4 Methodology

This section explores the pre-processing, feature extraction, feature fusion and training machine learning and deep learning models for classification of multimodal Hate Speech. The framework illustrating these methodologies is depicted in Figure 1. The design of the work is as follows:

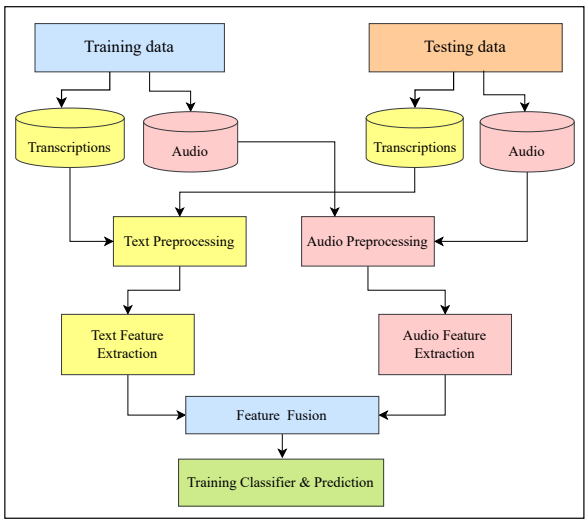


Figure 1: Overall flow diagram of the proposed framework

1. A methodology is implemented to extract audio and text features using Mel-Frequency Cepstral Coefficients (MFCCs) and XLM-RoBERTa respectively.

2. The extracted features from text and images are fused for further processing.

3. The fused features are used to train machine learning and deep learning models for classification of hate speech.

4.1 Pre-processing

Noises like stopwords, numbers, and punctuation that don't help with categorization are eliminated during pre-processing from the audio transcription. To eliminate Tamil, Malayalam, and Telugu stopwords, the github repositories'[1] Tamil, Malayalam, and Telugu are used.

4.2 Feature extraction

The pre-processed text data is then transformed into feature vectors using feature extraction techniques. This work utilizes XLM-RoBERTa to extract features from text using a transformer-based architecture designed for multilingual understanding. The process begins with tokenizing the text into subword units using Byte-Pair Encoding, ensuring effective handling of rare words and diverse languages. Each token is mapped to a high-dimensional feature that incorporates token, positional, and segment information. These features are passed through multiple transformer layers, where self-attention captures relationships between tokens, and feedforward networks refine the contextual representations. The hidden states generated in each layer provide rich contextualized features for each token. The final output is a set of dense feature vectors for Tamil, Malayalam and Telugu text.

To extract audio features, Mel-Frequency Cepstral Coefficients (MFCCs), is used for representation in audio processing. Each audio file was loaded in its original sampling rate using the Librosa library, ensuring the preservation of audio fidelity. MFCCs, capturing critical spectral characteristics of audio signals and modeled on human auditory perception, were computed with 13 coefficients per frame. These coefficients effectively summarize the power spectrum and frequency content of the audio signal. To derive a fixed-dimensional feature vector suitable for further analysis, we averaged the MFCCs over the time axis, reducing temporal variability while retaining essential features for Tamil, Malayalam, and Telugu audio. This approach ensures robust representation of audio characteristics.

[1] https://github.com/stopwords-iso/stopwords-iso

4.3 Feature Fusion & Classification

To integrate multimodal information from text and audio, we used a simple concatenation-based feature fusion strategy. The extracted text and audio features are combined along the feature dimension. This fused representation enables the model to leverage complementary insights from text and audio, improving its ability to capture multimodal context. Machine learning models like SVM, KNN, RF, NB and LR are trained on these fused features and is used for hate speech classification.

The Multimodal Classifier (MMC) integrates text and audio features for hate speech detection using a two-stage deep learning model. It consists of separate two-layer fully connected subnetworks for text and audio features, each utilizing ReLU activation, batch normalization, and dropout (0.1) for regularization. The extracted modality-specific features are concatenated and processed through a three-layer fusion network, which learns inter-modal relationships before classification using softmax activation. The model is trained for 50 epochs end-to-end with cross-entropy loss, using the Adam optimizer (learning rate = 6e-5) and batch size = 64. The hidden dimension is 256, ensuring effective feature representation and robust multimodal learning.

5 Results

Table 1 show the Accuracy (Acc), Precision (Pre), Recall (Rec), and F_1-score (F_1) achieved by various classifiers on the Tamil, Malayalam and Telugu datasets. This study evaluated both conventional machine learning and deep learning models to classify hate speech. The Multimodal Classifier (MMC) demonstrated superior performance compared to conventional machine learning models, including SVM, KNN, RF, NB and LR. The proposed Multimodal Classifier (MMC) achieved consistently higher accuracy of 72%, 78% and 37% and F_1-scores of 72%, 77% and 36% across the Tamil, Malayalam, and Telugu datasets, respec-

Table 1: Performance metrics of various models across Tamil, Malayalam and Telugu dataset.

	Tamil				Malayalam				Telugu			
Model	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
SVM	0.52	0.68	0.52	0.52	0.56	0.61	0.56	0.50	0.30	0.34	0.30	0.30
KNN	0.44	0.64	0.44	0.42	0.46	0.51	0.46	0.43	0.18	0.16	0.18	0.17
RF	0.32	0.60	0.32	0.26	0.46	0.36	0.46	0.35	0.28	0.24	0.28	0.21
NB	0.42	0.50	0.42	0.40	0.50	0.49	0.50	0.48	0.12	0.10	0.12	0.10
LR	0.28	0.30	0.28	0.21	0.42	0.46	0.42	0.38	0.16	0.13	0.16	0.14
MMC	0.72	0.75	0 .72	**0.72**	0.78	0.79	0.78	**0.77**	0.37	0.36	0.37	**0.36**

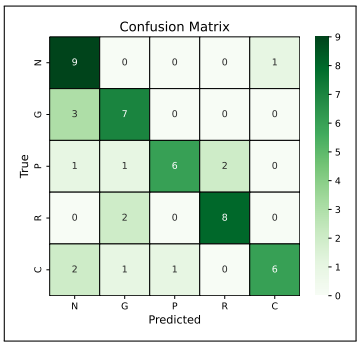


Figure 2: Confusion matrix of MMC for Tamil dataset.

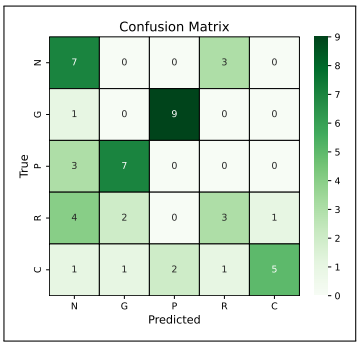


Figure 4: Confusion matrix of MMC for Telugu dataset.

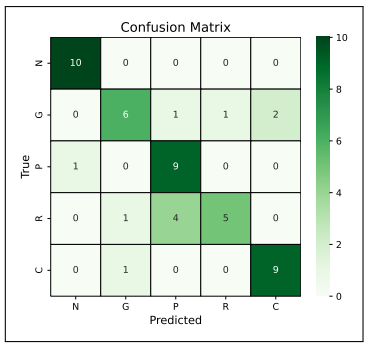


Figure 3: Confusion matrix of MMC for Malayalam dataset.

tively, effectively leveraging modality-specific features and their interactions. The confusion matrices for these models, illustrating their classification performance, are shown in Figures 2, 3, and 4, further highlighting the robustness of the MMC in handling complex multimodal data for hate speech detection tasks.

6 Conclusion

Hate speech is considered harmful as it can contribute to social divisions, violence, and harm to individuals dignity and rights. This study focused on detecting hate speech in Tamil, Malayalam, and Telugu languages using a multimodal approach that integrates text and audio features. The proposed Multimodal Classifier (MMC) outperformed traditional machine learning models such as SVM, KNN, RF, NB, and LR in terms of F_1-score across all three languages. MMC achieved an F_1-score of 72% in Tamil, 77% in Malayalam, and 36% in Telugu, significantly surpassing the performance of other models, which struggled to handle the complexities of code-mixed and linguistically diverse data. These results highlight the value of combining multiple modalities like text and audio to address the complexity of underrepresented languages. By demonstrating superior classification performance, this research emphasizes the importance of developing inclusive and robust hate speech detection systems, contributing to safer and more equitable digital spaces for diverse communities.

7 Limitations

While our multimodal approach improves hate speech detection in Tamil, Malayalam, and Telugu, certain limitations remain. Our model performs significantly better for Tamil and Malayalam but shows relatively lower performance for Telugu. This discrepancy may be due to smaller dataset

size, higher linguistic diversity, and the phonetic variations in Telugu, making it harder for the model to learn meaningful patterns. The dataset imbalance may also lead to biased predictions, affecting F_1-score for minority classes. The model struggles with dialect variations, code-mixed content, and informal social media language, which impacts its robustness in real-world scenarios. In the future, a stronger system can be developed by addressing these limitations.

The code for the proposed framework is available at:
https://github.com/Cshraddha153/A-Deep-Multimodal-Neural-Network-for-Hate-Speech-Detection-in-Dravidian-Languages.git

References

Selam Abitte Kanta, Grigori Sidorov, and Alexander Gelbukh. 2024. Selam@DravidianLangTech 2024:identifying hate speech and offensive language. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 91–95, St. Julian's, Malta. Association for Computational Linguistics.

Snehaan Bhawal, Pradeep Roy, and Abhinav Kumar. 2021. Hate speech and offensive language identification on multilingual code mixed text using bert. In *FIRE (Working Notes)*, pages 615–624.

Shankar Biradar, Sunil Saumya, Abhinav Kumar, and Ashish Singh. 2022. Pradvis vac: A socio-demographic dataset for determining the level of hatred severity in a low-resource hinglish language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md. Golam Rabiul Alam. 2021. Multi-modal hate speech detection using machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, page 4496–4499. IEEE.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Jose A. Diaz-Garcia and Joao Paulo Carvalho. 2025. A survey of textual cyber abuse detection using cutting-edge language models and large language models. *arXiv preprint arXiv:2501.05443*.

David W Hosmer and Stanley Lemeshow. 1989. *Applied Logistic Regression*. John Wiley & Sons.

George H John and Pat Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

Pallabi Kakati and Devendra Dandotiya. 2024. Automatic detection of hate speech in code-mixed indian languages in twitter social media interaction using dconvblstm-muril ensemble method. *Social Network Analysis and Mining*, 14(1):108.

Abhinav Kumar, Pradeep Kumar Roy, and Sunil Saumya. 2021. An ensemble approach for hate and offensive language identification in english and indo-aryan languages. In *FIRE (Working Notes)*, pages 439–445.

Abhinav Kumar, Sunil Saumya, and Jyoti Prakash Singh. 2020. NITP-AI-NLP@ HASOC-Dravidian-CodeMix-FIRE2020: a machine learning approach to identify offensive languages from dravidian code-mixed text. In *FIRE (Working notes)*, pages 384–390.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ankit Kumar Mishra, Sunil Saumya, and Abhinav Kumar. 2020. IIIT_DWD@ HASOC 2020: identifying offensive content in indo-european languages. In *FIRE (working notes)*, pages 139–144.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

M. Raphel, B. Premjith, K. Sreelakshmi, and B. R. Chakravarthi. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:12345–12356.

Pradeep Kumar Roy and Abhinav Kumar. 2025. Ensuring safety in digital spaces: Detecting code-mixed hate speech in social media posts. *Data & Knowledge Engineering*, page 102409.

Chava Sai, Rangoori Kumar, Sunil Saumya, and Shankar Biradar. 2024. IIITDWD_SVC@DravidianLangTech-2024: Breaking language barriers; hate speech detection in Telugu-English code-mixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*,

pages 119–123, St. Julian's, Malta. Association for Computational Linguistics.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2021. Offensive language identification in dravidian code mixed social media text. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pages 36–45.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2024. Filtering offensive language from multilingual social media contents: A deep learning approach. *Engineering Applications of Artificial Intelligence*, 133:108159.

Deepawali Sharma, Tanusree Nath, Vedika Gupta, and Vivek Kumar. 2024. Hate speech detection research in south asian languages: A survey of tasks, datasets, and methods. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–25.

Gopendra Vikram Singh, Soumitra Ghosh, and Pushpak Bhattacharyya. 2025. Unmasking offensive content: A multimodal approach with emotional understanding. *Multimedia Tools and Applications*, 84(1):1–25.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024a. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024b. A feature fusion and detection approach using deep learning for sentimental analysis and offensive text detection from code-mix malayalam language. *Journal of Intelligent Information Systems*.

Lanqin Yuan and Marian-Andrei Rizoiu. 2025. Generalizing hate speech detection using multi-task learning: A case study of political public figures. *Computer Speech Language*, 89:101690.

Muhammad Zamir, Moein Tash, Zahra Ahani, Alexander Gelbukh, and Grigori Sidorov. 2024. Lidoma@DravidianLangTech 2024: Identifying hate speech in Telugu code-mixed: A BERT multilingual. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 101–106, St. Julian's, Malta. Association for Computational Linguistics.

MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL Based Transformer Models for Detection of Abusive Tamil and Malayalam Text Targeting Women on Social Media

Md Mizanur Rahman, Srijita Dhar,
Md Mehedi Hasan, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1904116@student.cuet.ac.bd, dsrijita2001@gmail.com,
u1904067@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Social media has evolved into an excellent platform for presenting ideas, viewpoints, and experiences in modern society. But this large domain has also brought some alarming problems including internet misuse. Targeted specifically at certain groups like women, abusive language is pervasive on social media. The task is always difficult to detect abusive text for low-resource languages like Tamil, Malayalam, and other Dravidian languages. This paper presents a novel approach to detecting abusive Tamil and Malayalam texts targeting social media. A shared task on 'Abusive Tamil and Malayalam Text Targeting Women on Social Media Detection' has been organized by DravidianLangTech at NAACL-2025. We have implemented our model with different transformer-based models like XLM-R, MuRIL, IndicBERT, and mBERT transformers and the Ensemble method with SVM and Random Forest for training. We selected XLM-RoBERT for Tamil text and MuRIL for Malayalam text due to their superior performance compared to other models. After developing our model, we tested and evaluated it on the DravidianLangTech@NAACL 2025 shared task dataset. We found that XLM-R achieved the highest F1 score of 0.7873 on the test set, ranking 2nd among all participants for abusive Tamil text detections. In contrast, MuRIL had the highest F1 score of 0.6812 for abusive Malayalam text detections, ranking 10 th among all participants.

1 Introduction

The rise of social media has brought many benefits. Meanwhile, the prevalence of abusive text on social media has significantly increased in recent times. Abusive texting can result in cyberbullying, internet harassment, and other horrible acts. These kind words not only harm but also create a panicked and negative atmosphere for general users (Chen et al., 2017).

Abusive text detection is a critical task in social media content, especially when targeting vulnerable groups like children and women (Barker and Jurasz, 2021). There has been limited research on low-resource languages like Dravidian languages (Lee et al., 2018). Tamil and Malayalam are the two important languages in South India. In this paper, we address the abusive Tamil and Malayalam text detection. The cultural and linguistic nuances of these languages present unique challenges for natural language processing (NLP) tasks. So existing multilingual models often struggle to capture these nuances.

Our primary object of this paper is to detect abusive Tamil and Malayalam text targeting women on social media. We used different kinds of transformer models like XLM-R, MuRIL, m-BERT, and IndicBERT and ensemble methods with Random Forest, and SVM to train our model. The XLM-R and MuRIL transformer-based models were chosen because they outperformed Tamil and Malayalam texts, respectively.

The core contributions of our research work are as follows:-

1. We have developed an effective model to detect abusive Tamil and Malayalam text targeting women on social media with good generalizations.

2. We have conducted a systematic evaluation of multilingual transformer models and ensemble techniques. Comparing models for low-resource languages shows both their strengths and limitations.

The implementation details have been provided in the following GitHub repository:- https://github.com/Mizan116/DravidianLangTech-NAACL-2025.

2 Related Work

Abusive language detection identifies and filters damaging, insulting, or degrading information, especially on social media. Hate speech, sexism, homophobia, racism, bullying, and other verbal abuse are detected. Previously, there have been three main approaches to categorizing when it comes to studying how to identify abusive material, such as Tamil and Malayalam texts directed at women on social media.

Initially, the discipline was dominated by classical machine learning algorithms such as Support Vector Machines (SVMs) and Naïve Bayes and Random Forests for text mining techniques in Chen et al. (2017). Using manual feature engineering, abusive content detection can be achieved by extracting meaningful features such as n-grams and sentiment analysis, not handling code mixed language.

Because abusive material is complex and context-dependent. With the rise of deep learning, more powerful methods like LSTMs, especially CNNs and RNNs, are used in Founta et al. (2019), showing how Word2Vec, GloVe, and FastText which enhances the representation of abusive content, outperforming traditional feature engineering for abuse detection tasks. And findings depend on vast annotated datasets and difficulty in sarcasm and language.

Barker and Jurasz (2021) addressed the impact of text-based abuse on women on social media and the need for legislative frameworks to combat online violence, but not gender-targeted harassment in Tamil and Malayalam. Using supervised models and neural networks, The authors of Arellano et al. (2022) developed methods to recognize violent and aggressive material in Spanish, which can be applied to different languages and cultural contexts. Lexicon-based approach used in Lee et al. (2018) to enhance abusive(e.g., offensive terms, slurs) and non-abusive(e.g., polite terms or contextually mitigating words) word lists as the primary tool for detecting abusive text.

Transformer models, and neural networks, monitor relationships in sequential data like this phrase to learn context and meaning. It is a self-attention mechanism that replaces conventional RNNs and CNNs in order to capture long-range dependencies in sequences, setting the stage for models like BERT, GPT, and XLM-R, which are widely used for abusive content detection (Vaswani, 2017). Difficulties in identifying abusive comments in non-English languages, with a particular emphasis on Tamil and Malayalam.

3 Dataset and Task Overview

The DravidianLangTech 2023 (Priyadharshini et al., 2023) shared objective is to identify abusive comments in Tamil and Telugu (Priyadharshini et al., 2022) for only Tamil, including code-mixed and transliterated language, using models such as classical machine learning, deep learning, and transformers. It classified content as general abuse, which encompassed xenophobia, homophobia, and misogyny. Both of these papers do not explicitly mention the advancement of gender-specific abuse detection; rather, they concentrate on the fine-tuning of transformers for the detection of abusive comments.

We have utilized the abusive detection dataset from the DravidianLangTech@NAACL 2025 shared task, which includes two categories: Abusive and Non-abusive for both Tamil and Malayalam languages (Rajiakodi et al., 2025). The dataset is divided into three parts: training, development, and test. The test set does not have labels. Our models predicted the labels for that set. Table 1 provides an overview of the dataset, highlighting inconsistencies in label assignments. Some samples are tagged as 'Abusive', while others appear as 'abusive'. To maintain uniformity, we standardized all labels to 'Abusive' across the dataset.

2*Language	2*Split	Total Samples	2*Abusive	2*Non-Abusive
3*Tamil	Train	2790	1366	1424
	Dev	598	278	320
	Test	598	-	-
3*Malayalam	Train	2933	1531	1402
	Dev	629	303	326
	Test	629	-	-

Table 1: Category-wise distribution in the dataset

4 Methodology

This section provides an overview of the methodologies and approaches utilized to build the system using the previously described dataset and different transformer models. Methodology of our work is shown in Figure 1.

4.1 Preprocessing

The dataset is evaluated to determine its distribution and structure. The labels are encoded as Abusive: 1, Non-abusive: 0.

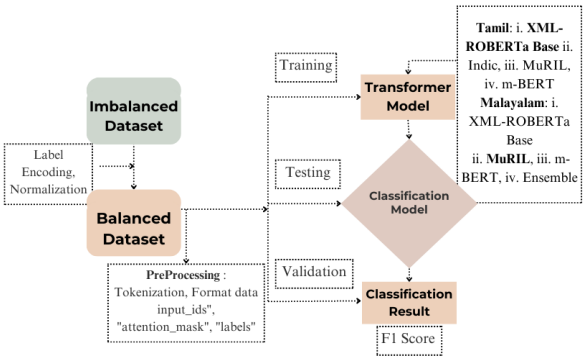


Figure 1: Methodology of our work

Model	lr	bs	ep	wd
XLM-R	2e-5	32	5	0.01
MuRIL	3e-5	32	7	0.1
Indic	2e-5	32	5	-
m-BERT	2e-5	32	5	-
Ensemble	3e-5	32	10	0.1

Table 2: Parameter setting in different model

Label normalization involves transforming variants of the abusive label to a consistent representation (e.g., 'abusive' is mapped to 'Abusive'). The normalization function applies to each sample in the dataset, ensuring label uniformity before encoding. Once standardized, labels are transformed into numerical representations for usage by machine learning models. The dataset is divided into training and validation sets using an 80%-20% ratio.

4.2 Model Selection and Training

We used XLM-RoBERTa, MuRIL, IndicBERT, and mBERT for Tamil, while for Malayalam, we employ XLM-RoBERTa, MuRIL, mBERT, and an ensemble approach. Tokenization is performed using the respective model tokenizers, ensuring uniform sequence lengths through padding and truncation. The dataset is converted into the Hugging Face Dataset format and preprocessed for PyTorch compatibility. During training, the model undergoes regular evaluations, and the best-performing model is saved. To improve efficiency, mixed precision (fp16) training is utilized. A data collator dynamically pads sequences, and evaluation metrics such as accuracy are computed to assess model performance. Early stopping is implemented to prevent overfitting by monitoring validation loss.

4.3 Evaluation and Testing

During model evaluation, we assessed performance using the new dataset for development to fine-tune hyperparameters and ensure optimal performance. Once the model had achieved satisfactory results, we proceeded with the test dataset for the final classification.

We have utilized a new test dataset, which contains unlabeled comments, is used to classify abusive and non-abusive comments. The trained model predicts the labels, distinguishing between abusive and non-abusive content. This ensured the model's ability to generalize effectively to unseen data.

5 Result and Analysis

In this section, we compare the results and analysis the different transformers performance based on some evaluation metrics. The performance of the various methods on the test set is showed in Table 2. The macro F1-score measures the supremacy of the models. However, we also consider other measures such as validation accuracy (Accuracy) and validation loss (Loss) also.

5.1 Parameter Setting

In Table 2, *lr*, *bs*, *ep*, and *wd* represents *learning_rate*, *batch_size*,*epochs*, *weight_decay* respectively. We have tuned different hyperparameters for finding the best model for the corresponding transformers.

5.2 Comparative Analysis

We have found that XLM-R has achieved the highest accuracy with 79% and an F1-score of 0.79 on the validation set for abusive Tamil text detection. For the Malayalam language, the MuRIL based model has given the highest output with 73% accuracy and an F1 score of 0.73 on the validation set. However, We have trained different transformer-based models and an ensemble method incorporating Random Forest (RF), Support Vector Machine (SVM), and MuRIL transformer altogether. However, we did not get optimal output from the ensemble methods and indic-BERT models. The model comparisons are shown in Table 3. XLM-R, MuRIL has given the close output in some cases when hyperparameters are tuned. So, after-all the MuRIL based model works finely for Malayalam whereas XML-R works better for Tamil text.

5.3 Metrics Evaluation

The performance of different model are evaluated by various metrics such as F1-score, Accuracy, Pre-

Tamil Text			
Transformer	**Loss**	**Accuracy**	**F1 Score**
XLM-R	0.3366	79%	**0.79**
MuRIL	0.6933	51%	0.51
Indic	0.9873	44%	0.42
m-BERT	0.8215	53%	0.51

Malayalam Text			
Transformer	**Loss**	**Accuracy**	**F1 Score**
XLM-R	0.4123	61%	0.60
MuRIL	0.3881	73%	**0.73**
m-BERT	0.8940	52%	0.51
Ensemble	0.6928	57%	0.56

Table 3: Comparison of different transformer models and ensemble methods

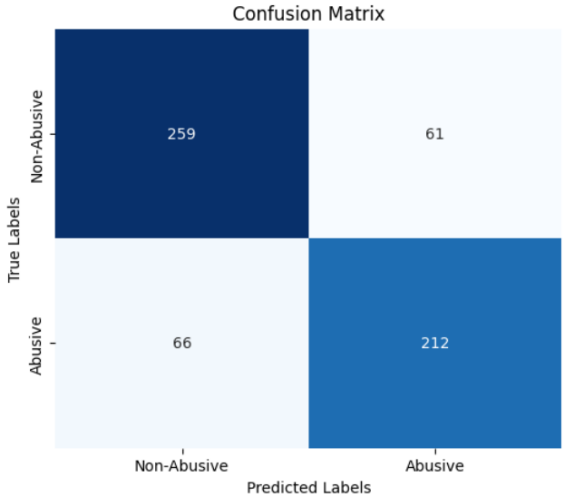


Figure 2: Confusion matrics of XLM-R transformer model for Tamil text

cision, and Recall on the test (Provided dev dataset) set. Figure 3 and Figure 3 show the confusion matrics of XLM-R and Muril based models for Tamil and Malayalam text respectively.

5.4 Error Analysis

Table 4 shows that the XLM-R based model performs well for abusive Tamil text detection and the MuRIL based model for Malayalam. In Table 4, *A*, *P*, *R*, and *F1* represents *Accuracy*, *Precision*, *Recall*, *F1_score* respectively. We have tuned different hyperparameters for finding the best model for the corresponding transformers. But in some cases, the non-abusive text is misclassified as abusive and vice versa. The confusion metrics show them well. These are due to the language morphology and lexical ambiguity, sarcasm, and irony

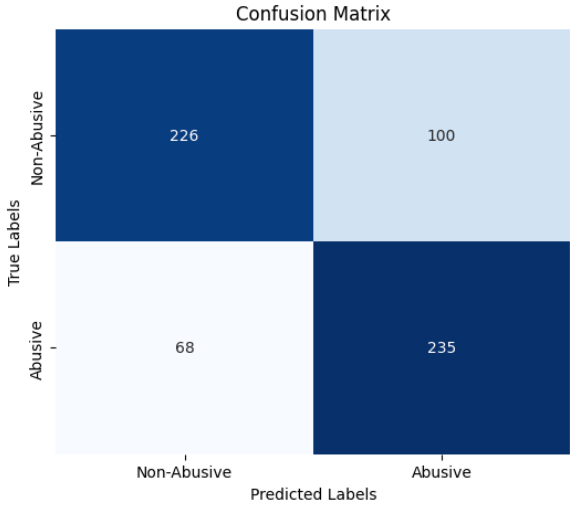


Figure 3: Confusion matrics of MuRIL transformer model for Malayalam text

Lang.	Model	A	P	R	F1
Tamil	XLM-R	79%	0.79	0.79	0.79
Malayalam	MuRIL	73%	0.74	0.73	0.73

Table 4: Evaluation metrics of our best models

when the context of the sentence is ambiguous. Rare words or Dialects may also be another reason for these misclassifications. The evaluation metrics of our best model for corresponding languages are shown in Table 4. Incorporating additional context using hierarchical models could help in better understanding the context. Fine-tuning multilingual transformers in domain-specific corpora may also improve performance.

6 Conclusion

In this research work, we have conducted a comparative study among different types of multilingual transformer and ensemble based techniques for abusive text detection in Tamil and Malayalam-two Dravidian languages. During the training and evaluation of the model, we used the Dravidian-LangTech provided annotated datasets. Although we try to use different transformers and ensemble methods with RF and SVM, MuRIL and XLM-R has given the better output comparing others. Surprisingly, ensemble techniques and Indic-BERT has performed poorly. However, we have tuned some hyperparameters for the model and got decent outputs.

Limitations

While our approach demonstrates better performance, it has certain limitations also. First of all, the provided dataset is quite small. The impact of the dataset on model development is visible in the result and error analysis section. Secondly, our model shows limitations in capturing the sarcasm, irony, or implicit abusive content. As these are low resources languages and due to their native morphology, capturing the context is challenging.

References

Luis Joaquín Arellano, Hugo Jair Escalante, Luis Villaseñor Pineda, Manuel Montes y Gómez, and Fernando Sanchez-Vega. 2022. Overview of da-vincis at iberlef 2022: Detection of aggressive and violent incidents from social media in spanish.

Kim Barker and Olga Jurasz. 2021. Text-based (sexual) abuse and online violence against women: Toward law reform? In *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*, pages 247–264. Emerald Publishing Limited.

Hao Chen, Susan Mckeever, and Sarah Jane Delany. 2017. Harnessing the power of text mining for the detection of abusive content in social media. In *Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK*, pages 187–205. Springer.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114.

Ho-Suk Lee, Hong-Rae Lee, Jun-U Park, and Yo-Sub Han. 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, 113:22–31.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and booktitle = Kumaresan, Prasanna Kumar". 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

MNLP@DravidianLangTech 2025: Transformer-based Multimodal Framework for Misogyny Meme Detection

Shraddha Chauhan
Department of ECE
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
shraddha.20224147@mnnit.ac.in

Abhinav Kumar
Department of CSE
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
abhik@mnnit.ac.in

Abstract

A meme is essentially an artefact of content-usually an amalgamation of a image and text that spreads like wildfire on the internet, usually shared for amusement, cultural expression, or commentary. They are very much similar to an inside joke or a cultural snapshot that reflects shared ideas, emotions, or social commentary, remodulated and reformed by communities. Some of them carry harmful content, such as misogyny. A misogynistic meme is social commentary that espouses negative stereotypes, prejudice, or hatred against women. The detection and addressing of such content will help make the online space inclusive and respectful. The work focuses on developing a multimodal approach for categorizing misogynistic and non-misogynistic memes through the use of pretrained XLM-RoBERTa to draw text features and Vision Transformer to draw image features. The combination of both text and images features are processed into a machine learning and deep learning model which have attained F_1-scores 0.77 and 0.88, respectively Tamil and Malayalam for misogynist Meme Dataset.

1 Introduction

Memes have become one of the most powerful ways of expressing on social media, often full of humor, satire, and cultural commentary. Seemingly innocuous in nature, memes can be a means of disseminating harmful content (Weber et al., 2020) that enforces gender biases and stereotypes. Thus, it is crucial to identify such content in order to bring about a more respectful and inclusive digital space (Rao and Kalyani, 2022; Kumar et al., 2021). The classification of misogyny in Tamil and Malayalam memes faces different types of challenges because of linguistic as well as cultural characteristics of these languages (Fersini et al., 2022). As Dravidian languages are low resource languages which possess unique syntactic structures and vocabulary and idiomatic expressions (Singh et al., 2025). There is a scarcity of labeled datasets within these languages, which complicates the training of Machine and Deep learning models. Lack of datasets makes it hard to develop effective machine learning models because they require enough good-quality and labelled data to learn complex patterns of misogyny in social media memes.

Adding text and images to memes makes it even more complex as its often the use of visual elements to relay context and meaning which might be hard for text-only models to understand. The multimodal approach is necessitated to capture the complexity between the textual and visual components of memes (H et al., 2024). With deep learning models such as transformers in text and image feature extraction, along with feature fusion strategies, it is possible to raise the classification accuracy of misogynistic memes. Therefore, this work proposes a multimodal deep learning-based model for the identification of misogynist memes.

The rest of the paper is arranged as follows: Section 2 contains related work, Section 3 discusses the dataset & task, and Section 4 presents the proposed methodology. In Section 5, the outcome of the proposed model is listed and concluded in Section 6, and Section 7 details the limitations of the proposed framework.

2 Related Work

Detection of misogynistic content in memes in low-resource languages has started to gain greater attention in the recent past (Kumar et al., 2021; Ponnusamy et al., 2024). Memes that convey meaning using textual and visual elements introduce an interesting challenge (Priyadharshini et al., 2022). In recent years, several studies focused on multimodal approaches that have combined text as well as image data for classifying hate speech and misogynistic content. Recent studies have adapted techniques like data augmentation or transfer learning

from related languages to improve the models' performance on those low-resource languages (Joshi et al., 2020). For the detection of misogyny in meme analysis, datasets play an important role. A prominent MIMIC dataset (Singh et al., 2024) focuses on low-resource Hindi-English code-mixed language, thus allowing the detection of misogyny. A study by (Rizzi et al., 2023) explores various approaches in detecting misogynistic content in memes. They compares different approaches in the integration of textual and visual data, one unimodal and the other multimodal approach, respectively.

Multimodal approaches (Jindal et al., 2024) are vital while considering memes since they are based on the visual component and textual information. Convolutional Neural Networks were traditionally applied to extract features from images. The recent alternative for image processing has been the Vision Transformers (Dosovitskiy et al., 2014). Deep learning-based techniques are used in various studies like (Garcia et al., 2021), where the use of textual and visual information enhances performance. The study (Kiela et al., 2019) proved the multimodal embeddings effectiveness in cross-modal retrieval and classification tasks, where textual and visual features are combined to improve the understanding of meme content. Misogyny in memes is challenging due to the complexities present in meme content, which often involves humor, cultural references, and contextual elements (Chakravarthi et al., 2024). Traditional machine learning models struggle to capture these complexity. Study by (Suryawanshi et al., 2020) and (Beigi et al., 2020) addresses these challenges by incorporating both image and text features for detecting hate speech. However, these approaches face limitations due to the paucity of labeled data in Tamil and Malayalam, and labeled datasets are often small and imbalanced.

Data imbalance is a massive concern for various tasks and becomes more acute in the context of misogynistic meme detection (Hossain et al., 2022; Gasparini et al., 2022). Several methods have been proposed to address such imbalance, such as oversampling the minority class or applying cost-sensitive learning techniques (Buda et al., 2018). Synthetic data generation and semi-supervised learning have been explored in order to enhance classifier performance in cases where the training data is limited (Zhang et al., 2020). These techniques can be specifically helpful when used with meme datasets in low-resource languages, mitigating the imbalance problem of the data.

3 Dataset & Task

The Tamil dataset distribution consists of 1,133 memes for training and 356 memes for testing purposes and Malayalam dataset consists of 640 training and 200 test data are shown in Table 1. Each meme in the dataset contains both pictorial content and overlaid text (Chakravarthi et al., 2025). The memes are classified based on the presence or absence of misogyny. However, the dataset is imbalanced, with a significantly higher number of non-misogynistic memes compared to misogynistic ones for both Tamil and Malayalam datasets.

Table 1: Datasets and their distribution.

Dataset	Label	Train	Val	Test
Tamil	Misogynistic	285	74	89
	Non-Misogynistic	848	210	267
Total		1,133	284	356
Malayalam	Misogynistic	259	63	78
	Non-Misogynistic	381	96	122
Total		640	159	200

Figure 1: Examples of misogyny detection in memes from Tamil and Malayalam datasets.

A sample memes from the Tamil and Malayalam can be seen in Figure 1 with their transcription and label.

4 Methodology

This section explores the pre-processing, feature extraction, feature fusion and training machine learning and deep learning models for classification of misogyny memes. The framework illustrating these methodologies is depicted in Figure 2. The design of the work is as follows: (i) Extract image and text

features using transfer learning with the pre-trained ViT-Base-Patch16-224 and XLM-RoBERTa, respectively. (ii) Fuse the extracted text and image features for further processing. (iii) Train machine learning and deep learning models on the fused features to classify misogynistic content.

4.1 Pre-processing

The text extracted from the transcription of memes consists of noise, such as stopwords, digits, and punctuation which do not contribute to the classification. English stopwords available at NLTK library, Tamil and Malayalam stopwords available at github repositories[1] are used as references to remove English, Tamil and Malayalam stopwords, respectively. Vision Transformer preprocesses Tamil and Malayalam memes by resizing the images to a fixed input size of (224×224) pixels.

4.2 Feature extraction

The pre-processed text data is transformed into feature vectors using feature extraction techniques. This work utilizes XLM-RoBERTa to extracts features from text. The process begins with tokenizing the text into subword units using Byte-Pair Encoding, ensuring effective handling of rare words and diverse languages. Special tokens like <s> and </s> are added to structure the sequence, and each token is mapped to a high-dimensional embedding that incorporates token, positional, and segment information. These embeddings are passed through multiple transformer layers, where self-attention captures relationships between tokens, and feedforward networks refine the contextual representations. Hidden states generated at each layer provide rich, contextualized embeddings for each token, while the state of the <s> token often serves as a global representation for the input. The final output is a set of dense feature vectors for Tamil and Malayalam text.

The resized images are split into non-overlapping 16x16 pixel patches, flattened, and projected into a high-dimensional space using a linear layer, creating patch embeddings. These embeddings, combined with positional encodings, are passed through multiple self-attention layers of the Vision Transformer, where relationships between patches are learned. This process captures both local and global contextual features from the meme, enabling effective representation for classification task.

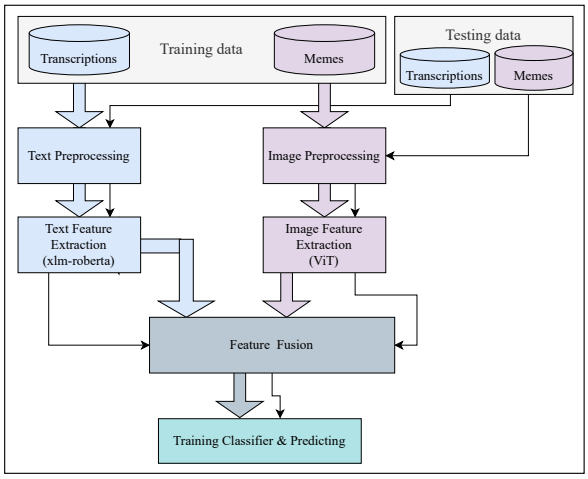


Figure 2: Block diagram of experimental work.

4.3 Feature Fusion & Classification

To fuse text and image features, we employ a simple concatenation-based feature fusion strategy. This fused representation allows the model to utilize both textual and visual modalities. To classify misogynistic memes, we utilized machine learning models and deep learning architectures. The machine learning models included K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB), which were trained using the fused feature embeddings derived from text and image modalities. For deep learning-based classification, we employed Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks to capture temporal dependencies in the fused features.

The Multimodal Classifier (MMC) integrates text and audio features for hate speech detection using a two-stage deep learning model. It consists of separate two-layer fully connected subnetworks for text and image features, each utilizing ReLU activation, batch normalization, and dropout (0.3) for regularization. The extracted modality-specific features are concatenated and processed through a three-layer fusion network, which learns inter-modal relationships before classification using softmax activation. The model is trained for 180 epochs end-to-end with binary cross-entropy loss, using the Adam optimizer (learning rate = 5e-5) and batch size = 32. The hidden dimension is 256, ensuring effective feature representation and robust multimodal learning.

[1] https://github.com/stopwords-iso/stopwords-iso

5 Results

Tables 2 and 3 show the Accuracy (Acc), Precision (Pre), Recall (Rec), and F_1-score (F_1) achieved by various classifiers on the Tamil and Malayalam datasets, respectively. The machine learning models considered include KNN, SVM, RF, and NB. For deep learning approaches, LSTM, GRU, and MMC were used.

Table 2: Performance of classifiers on fused embeddings for Tamil data

Classifier	Acc	Pre	Rec	F1
KNN	0.85	0.86	0.71	0.75
SVM	0.83	0.80	0.72	0.75
RF	0.83	0.88	0.66	0.69
NB	0.72	0.69	0.74	0.69
LSTM	0.83	0.77	0.75	0.76
GRU	0.81	0.75	0.71	0.73
MMC	0.81	0.75	0.79	**0.77**

Table 3: Performance of classifiers on fused embeddings for Malayalam data

Classifier	Acc	Pre	Rec	F1
KNN	0.85	0.90	0.81	0.83
SVM	0.85	0.85	0.85	0.85
RF	0.85	0.86	0.83	0.84
NB	0.81	0.81	0.81	0.81
LSTM	0.88	0.87	0.87	0.87
GRU	0.88	0.87	0.88	**0.88**
MMC	0.86	0.86	0.86	0.86

Among the models, the MMC outperformed others on the Tamil dataset, achieving an F_1-score of 77%, demonstrating its capability to process both textual and visual features effectively. On the Malayalam dataset, the GRU achieved the highest performance, with an F_1-score of 88%, indicating its superior ability to handle textual intricacies in this low-resource language whereas proposed MMC model achieved comparable performance with F_1-score of 0.86. Figures 3 and 4 illustrate the confusion matrices for the best-performing classifiers.

6 Conclusion

Misogyny refers to the discrimination against women, has been a persistent issue in society, affecting both offline and online spaces. There is increasing concern about its appearance in digital media. We apply transformer-based models to detect misogynistic content in a code-mixed Tamil and Malayalam. The results were encouraging: with an F_1 score of 77% for Tamil and an F_1 score

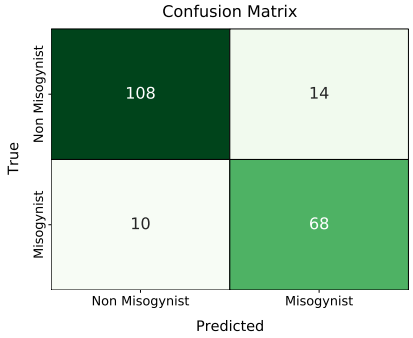


Figure 3: Confusion matrix of MMC for Tamil dataset.

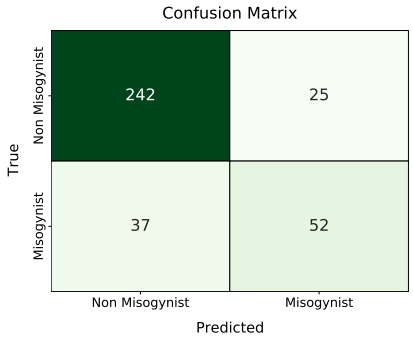


Figure 4: Confusion matrix of GRU for Malayalam dataset.

of 88% for the Malayalam. This indicates that the model was successful in detecting misogynistic content, even where code-mixing was problematic and training data were limited. These results illustrate the capabilities of multimodal deep learning methods to understand and identify misogyny in low resource languages.

7 Limitations

Despite achieving strong performance in misogynistic meme classification for Tamil and Malayalam, our approach has certain limitations. First, the dataset size may not be sufficient to generalize across all variations of misogynistic memes, particularly those with subtle, implicit biases or sarcasm, which can be difficult for models to detect. The model also struggles with code-mixed content, low-resolution images, and heavily stylized fonts, which can distort both textual and visual understanding. While Malayalam models perform better than Tamil models, this discrepancy could stem from dataset composition, linguistic variations, or image-text alignment differences. In the future, a stronger system can be developed by addressing these limitations.

The code for the proposed framework is

available at:
https://github.com/Cshraddha153/Transformer-based-Multimodal-Framework-for-Misogyny-Meme-Detection.git

References

Gita Beigi, Haiyong Ho, Kristina Lerman, and Benjamin C. Wallace. 2020. Hate speech detection in memes: A survey of approaches, datasets, and challenges. In *Proceedings of the 2020 IEEE/ACM International Conference on Web Search and Data Mining (WSDM)*, pages 322–330.

Maciej Buda, Alan Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. In *Proceedings of the International Conference on Computational Intelligence and Neuroscience*, pages 1–6.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Alexey Dosovitskiy, Philipp Fischer, Jörg R Springenberg, Martin Riedmiller, and Marc Brockschmidt. 2014. Discriminative unsupervised feature learning with exemplar convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 764–772.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Vanessa Garcia, Yi Chang, Yifan Xu, and Enrique Alfonseca. 2021. Multimodal meme classification: Leveraging textual and visual features for robust meme detection. In *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*, pages 3276–3280.

I. Gasparini, A. Singh, G. Vasan, A. Narayan, and A. Deshmukh. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. In *Proceedings of the 2022 IEEE International Conference on Data Mining (ICDM 2022)*, pages 1267–1274, Chennai, India. IEEE.

Shaun H, Samyuktaa Sivakumar, Rohan R, Nikilesh Jayaguptha, and Durairaj Thenmozhi. 2024. Quartet@LT-EDI 2024: A SVM-ResNet50 approach for multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 221–226, St. Julian's, Malta. Association for Computational Linguistics.

M. Hossain, M. Islam, M. Rahman, and S. Shahin. 2022. Memosen: A multimodal dataset for meme sentiment analysis in bengali. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*, pages 1651–1658, Marseille, France. European Language Resources Association.

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Douwe Kiela, Mohamed Elhoseiny, Lin Zhang, Marco Baroni, and Geoffrey Hinton. 2019. Supervised multimodal hashing for scalable cross-modal retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1740–1751.

Abhinav Kumar, Pradeep Kumar Roy, and Jyoti Prakash Singh. 2021. A deep learning approach for identification of arabic misogyny from tweets. In *FIRE (Working Notes)*, pages 831–838.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

B Narasimha Rao and V Kalyani. 2022. A study on positive and negative effects of social media on society. *Journal of Science & Technology (JST)*, 7(10):46–54.

Giorgia Rizzi, Wei Zhang, and Martin Hall. 2023. Detecting misogyny in memes: A comparative study of unimodal and multimodal approaches. In *Proceedings of the International Conference on Multimodal Analysis*, pages 199–208.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2025. Misogynistic attitude detection in youtube comments and replies: A high-quality dataset and algorithmic models. *Computer Speech Language*, 89:101682.

Ravi Singh, Aman Sharma, and Vikash Gupta. 2024. Misogyny identification in multimodal internet content (mimic). *Journal of Social Media Studies*, 12(4):345–368.

Amit Suryawanshi, Sachin Jadhav, and Venkatesh Rajendran. 2020. Hate speech detection in memes: A comparative analysis of image and text-based approaches. In *Proceedings of the 2020 International Conference on Computational Intelligence and Data Science (ICCIDS)*, pages 287–292.

Mathias Weber, Christina Viehmann, Marc Ziegele, and Christian Schemer. 2020. Online hate does not stay online–how implicit and explicit attitudes mediate the effect of civil negativity and hate in user comments on prosocial behavior. *Computers in human behavior*, 104:106192.

Yu Zhang, Xue Xu, Hongzhi Yang, Zhenyu Yu, and Xiang Yu. 2020. An overview of deep learning-based approaches for multimodal hate speech detection. *Journal of Artificial Intelligence Research*, 69:545–576.

Code_Conquerors@DravidianLangTech 2025: Deep Learning Approach for Sentiment Analysis in Tamil and Tulu

Harish Vijay V, Ippatapu Venkata Srichandra, Pathange Omkareshwara Rao, Premjith B.
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham India
harishvijay0204@gmail.com, ippatapuvenkatasrichandra@gmail.com,
cb.en.u4aie22039@cb.students.amrita.edu, b_premjith@cb.amrita.edu

Abstract

In this paper we propose a novel approach to sentiment analysis in languages with mixed Dravidian codes, specifically Tamil-English and Tulu-English social media text. We introduce an innovative hybrid deep learning architecture that uniquely combines convolutional and recurrent neural networks to effectively capture both local patterns and long-term dependencies in code-mixed text. Our model addresses critical challenges in low-resource language processing through a comprehensive preprocessing pipeline and specialized handling of class imbalance and out-of-vocabulary words. Evaluated on a substantial dataset of social media comments, our approach achieved competitive macro F1 scores of 0.3357 for Tamil (**ranked 18**) and 0.3628 for Tulu (**ranked 13**).

Keywords: Sentiment Analysis, Code mixed text, Dravidian languages, Deep Learning, CNN BiLSTM, Tamil-English, Tulu-English.

1 Introduction

It is difficult to analyze sentiment in code-mixed text because of the intricacy of code-switching at many linguistic levels and the absence of annotated datasets. This is particularly relevant for Dravidian languages like Tamil and Tulu, where social media communications frequently mix local languages with English, often written in non native scripts. While recent advances in natural language processing have shown promising results for monolingual text, these systems typically perform poorly on code mixed content, highlighting the need for specialized approaches. Our research addresses these challenges through three main contributions: (1) a novel hybrid deep learning architecture combining CNN BiLSTM networks, specifically designed for code mixed text processing, (2) an effective preprocessing pipeline that handles the unique characteristics of Dravidian code mixed text, and (3) a systematic approach to addressing class imbalance through weighted learning, validated on a diverse dataset comprising over 20,000 Tamil English and 13,000 Tulu English social media comments. **The code is available in this Github Code link.**

2 Related Works

A study from the RANLP 2023 shared task Kanta and Sidorov (2023) et al. examined sentiment study in code mixed dataset. The motive of this study was to categorize YouTube comments into mixed emotions, Neutral, Negative and Positive. The dataset, collected from social media posts, included training, development, and test sets. They used SVM for classification, which got an macro f1 score of 0.147 for tamil-english and 0.518 for tulu-english.

One notable limitation was the low accuracy for tamil-english dataset. For a focused study Hegde et al. (2022) et al. on sentiment examination in data-scarce languages, researchers created a trilingual code mixed Tulu collections with 7,171 YouTube comments. This dataset addresses the lack of tagged data for Tulu. Baseline evaluations using machine learning models showed promising results, though challenges persist due to the informal structure of social media text.

A study by Kannadaguli (2021) et al. focused on Tulu English code mixed text and created the first platinum standard dataset for sentiment analysis. Machine learning and deep learning methods performed better than unsupervised approaches. A recent study by the MUCS team focused on Tamil and Tulu text classification. Using LinearSVC and an ensemble of five classifiers, the team trained models on features derived from word and character n grams.Prathvi et al. (2024).

In a study by Ehsan et al. (2023), sentiment study of code mixed Tulu and Tamil YouTube reviews was tackled using Bidirectional LSTM networks.

The models utilized ELMo embeddings fed and trained using larger unannotated code mixed collections for better contextual understanding. Another recent study Tripty et al. (2024). focuses on sentiment study for Tamil and Tulu code mixed text using transformer based models. It found that mBERT and XLM R outperformed others.

This study Chakravarthi et al. (2021). created the multimodal sentiment study dataset for Tamil and Malayalam. They collected YouTube review videos, generated captions, and labeled them for sentiment. The inter annotator consent was verified using Fleiss's Kappa.

Another research by Ponnusamy et al. (2023) tackled sentiment detection in code-mixed social platform comments, which often mix scripts and deviate from grammatical rules. This was achieved using preprocessing and feature extraction techniques along with logistic regression models.

In a recent study, Shetty (2023) et al. overcame the hurdle of sentiment analysis in the tulu dataset. Previous works on code-mixed text demonstrated the effectiveness of machine learning and transformer based models in handling script mixing and linguistic diversity.

However, class imbalance and a lack of annotated datasets remain critical issues for low resource languages. To mitigate these challenges, a new annotated corpus for Tulu was developed and evaluated using standard preprocessing and classification techniques, achieving encouraging results in sentiment classification. This advancement offers a promising foundation that can inspire further work and refinement in low-resource settings.

In another study by Rachana et al. (2023). in which they used fasttext vector representation to train machine learning model for Tulu and Tamil sentences. The models achieved F1-scores of 0.14 for Tamil and 0.204 for Tulu, indicating that there is room for further improvements. These findings underscore the potential benefits of exploring alternative feature representations and tuning strategies to push performance even further.

3 Task Details

Abeera et al. (2023) Social media messages require sentiment analysis because they are mainly code mixed data for dravidian languages. The dataset that we used were code-mixed data Tulu English and Tamil English dataset for the sentiment analysis. Chakravarthi et al. (2020).Lavanya et al. (2024)(Durairaj et al., 2025). The class imbalance issue in the presented dataset illustrates issues that arise in the actual world. The dataset description is displayed in table 1 and 2.

Labels	Train Data	Dev Data	Test Data
NotTulu	4400	543	474
Positive	3769	470	453
Neutral	3175	368	343
Mixed	1114	143	120
Negative	843	118	88

Table 1: Tulu dataset description.

Labels	Train Data	Dev Data	Test Data
Positive	18145	2272	1983
Unknown State	5164	619	593
Mixed Feelings	3662	472	425
Negative	4151	480	458

Table 2: Tamil dataset description.

4 Methodology

Figure 1 represents the proposed workflow, where we consider the raw text data as input. Various preprocessing steps are performed, followed by addressing the class imbalance problem. Later, the target variables will be encoded and later we trained a hybrid model and later we evaluated the model on various evaluation scores.

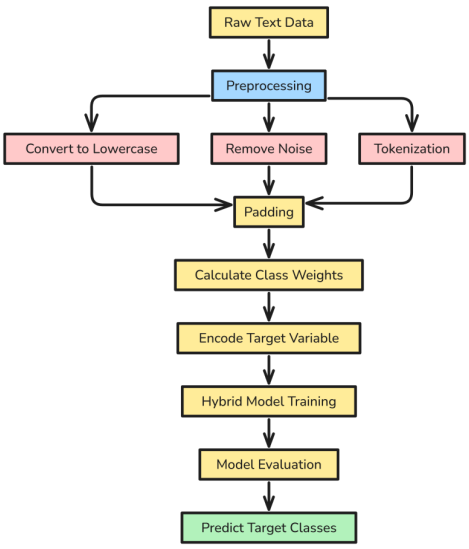


Figure 1: Proposed Methodology.

4.1 Data Preprocessing

In the Initial step, the raw text data was taken, and various preprocessing methods were performed on the text. A function was defined to convert all the text to lowercase. Noise that includes usernames (@), hashtags (#), and punctuation was filtered out from the text using regular expressions. Next, tokenization was performed on the individual words. To handle variable length inputs, each sentence was padded with zeros up to a maximum length to ensure a consistent input size for the model.

The number of unique words in the combined training and validation corpus was used to determine the vocabulary size. To handle words that weren't in the training data, a unique token called <OOV> was utilized.

We used the class weight method in order to address the issue of class inequality. The weight of each class was determined by dividing the total number of samples by the frequency of each class. The label encoding method was ultimately used to transform the target variable into numerical values. The explicit data from the task was utilized for testing and validation, and the dataset from the task was used to train the model.

4.2 Model Architecture

Convolutional and recurrent layers are combined in the design to effectively detect the dataset's long term dependencies as well as local patterns. The model begins with an embedding layer that turns every word into a 100-size dense vector. Word semantic associations can be captured by this layer.

The input sequences are then subjected to a 1D convolutional layer that uses 128 filters with a kernel size of 5 to extract local features. After applying the relu activation function, padding makes sure the output shape is the same as the input shape. To down sample the data, we added a max pooling layer with a pool size of 5.

A bidirectional LSTM layer Kumar et al. (2017). was incorporated to capture long term dependencies. After that, a dense layer using the ReLU activation function with 32 fully connected units comes next. Finally, the output layer, determined by the number of unique tokens, predicts the target classes. The table-3 and 4 tells about the model summary for the tamil and tulu datasets.

4.3 Hyperparameters Setting

The model was configured with the Adam optimizer set to a learning rate of 0.01, and categorical crossentropy was employed as the loss function, making it well-suited for this multi-class classification task. Training was conducted over 100 epochs with a batch size of 32.

Layer	Output Shape	Param
embedding	(None, 40, 100)	2,713,300
conv1d	(None, 40, 128)	64,128
maxpooling1d	(None, 8, 128)	0
bidirectional	(None, 64)	41,216
dense	(None, 32)	2,080
dense	(None, 5)	165

Table 3: Model summary for the tamil dataset.

Layer	Output Shape	Param
embedding	(None, 178, 100)	7,115,400
conv1d	(None, 178, 128)	64,128
maxpooling1d	(None, 36, 128)	0
bidirectional	(None, 64)	41,216
dense	(None, 32)	2,080
dense	(None, 4)	132

Table 4: Model summary for the tulu dataset.

5 Results

5.1 Performance Analysis

The proposed deep learning model was evaluation on differnet metrics which include accuracy, macro f1 score, macro precision score, and macro recall score.

Evaluation Metrics	Scores
Accuracy	0.5114
Macro precision	0.3624
Macro recall	0.3427
Macro f1	0.3357

Table 5: Evaluation scores of the Tamil dataset.

Tables 5 and 6 present the evaluation scores of the Tamil and Tulu datasets. From these tables, it

is observed that the Tulu dataset achieved a macro f1 Score of 0.3628, while the Tamil dataset accomplished a score of 0.3357. Figures 2 and 3 Depict the confusion matrices for the Tamil and Tulu datasets using the proposed model, showing the number of correct predictions along the diagonal across the classes.

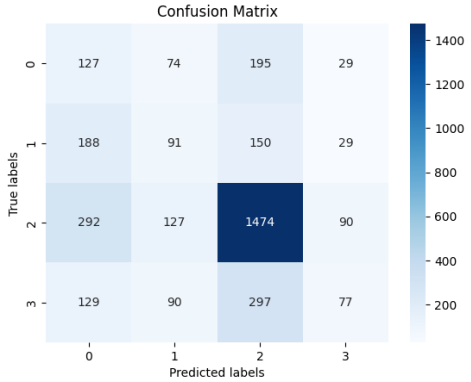


Figure 2: The Tamil dataset's confusion matrix.

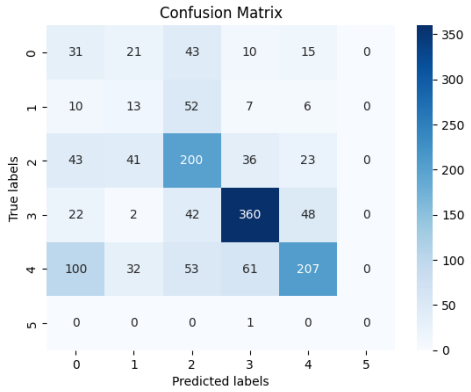


Figure 3: The Tulu dataset's confusion matrix.

Evaluation Metrics	Scores
Accuracy	0.5483
Macro precision	0.3721
Macro recall	0.3676
Macro f1	0.3628

Table 6: Evaluation scores of the Tulu dataset.

6 Conclusion

In this research paper we used tamil-english and tulu-english datasets in which using a combined convolutional and recurrent architecture so that it could capture both local and long term dependencies present in the text. From the results it was observed that for tamil dataset we achieved a macro f1 score of 0.3357 and 0.3628 for tulu dataset.

The OOV problem was addressed using a special <OOV> token, and class imbalance was mitigated through class weighting. Future research will explore the use of pre-trained multilingual models like mBERT and IndicBERT, hyperparameter tuning, and the addition of attention mechanisms to enhance performance. Additionally, comparisons with CNN and LSTM models will be conducted.

7 Limitations

The primary drawback of the suggested approach is the lack of an attention mechanism, which would have enabled the model to concentrate on the crucial segments of the input sequence. Additionally, models like IndicBERT and ModernBERT, which are trained on Indian-context data, could have provided better contextual understanding.

References

S. Kanta and G. Sidorov. Selam@ dravidianlangtech: Sentiment analysis of code mixed dravidian texts using svm classification. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 176–179, September 2023.

A. Hegde, M. D. Anusha, S. Coelho, H. L. Shashirekha, and B. R. Chakravarthi. Corpus creation for sentiment analysis in code mixed tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under Resourced Languages*, pages 33–40, June 2022.

P. Kannadaguli. A code diverse tulu english dataset for nlp based sentiment analysis applications. In *2021 Advanced Communication Technologies and Signal Processing (ACTS)*, pages 1–6. IEEE, December 2021.

B. Prathvi, K. Manavi, K. Subrahmanyapoojary, A. Hegde, G. Kavya, and H. Shashirekha. Mucs@ dravidianlangtech 2024: A grid search approach to explore sentiment analysis in code mixed tamil and tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 257–261, March 2024.

T. Ehsan, A. Tehseen, K. Sarveswaran, and A. Ali. Alphabrains@ dravidianlangtech: Sentiment analysis of code mixed tamil and tulu by training contextualized elmo word representations. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 152–159, September 2023.

Z. Tripty, M. Nafis, A. Chowdhury, J. Hossain, S. Ahsan, A. Das, and M. M. Hoque. Cuetsentimentsillies@

dravidianlangtech eacl2024: Transformer based approach for sentiment analysis in tamil and tulu code mixed texts. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 234–239, March 2024.

B. R. Chakravarthi, K. P. Soman, R. Ponnusamy, P. K. Kumaresan, K. P. Thamburaj, and J. P. McCrae. Dravidianmultimodality: A dataset for multi modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*, 2021.

K. K. Ponnusamy, C. Rajkumar, P. K. Kumaresan, E. Sherly, and R. Priyadharshini. Vel@ dravidianlangtech: Sentiment analysis of tamil and tulu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 211–216, September 2023.

P. Shetty. Poorvi@ dravidianlangtech: Sentiment analysis on code mixed tulu and tamil corpus. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 124–132, September 2023.

K. Rachana, M. Prajnashree, A. Hegde, and H. L. Shashirekha. Mucs@ dravidianlangtech2023: Sentiment analysis in code mixed tamil and tulu texts using fasttext. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 258–265, September 2023.

V. P. Abeera, S. Kumar, and K. P. Soman. Social media data analysis for malayalam youtube comments: Sentiment analysis and emotion detection using ml and dl models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51, September 2023.

B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, and J. P. McCrae. Corpus creation for sentiment analysis in code mixed tamil english text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under resourced Languages (SLTU) and Collaboration and Computing for Under Resourced Languages (CCURL)*, pages 202–210, Marseille, France, May 2020. Online available: https://aclanthology.org/2020.sltu-1.28.

S. K. Lavanya, A. Hegde, B. R. Chakravarthi, H. L. Shashirekha, R. Natarajan, S. Thavareesan, R. Sakuntharaj, T. Durairaj, P. K. Kumaresan, and C. Rajkumar. Overview of second shared task on sentiment analysis in code mixed tamil and tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta, March 2024.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, 2025.

S. S. Kumar, M. A. Kumar, and K. P. Soman. Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets. In *Mining Intelligence and Knowledge Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings 5*, pages 320–334. Springer International Publishing, 2017.

KEC_TECH_TITANS@DravidianLangTech 2025: Abusive Text Detection in Tamil and Malayalam Social Media Comments Using Machine Learning

Malliga Subramanian[1], Kogilavani S V[1], Deepiga P[1], Dharshini S[1], Ananthakumar S[1], Praveenkumar C[1]

[1]*Kongu Engineering College, Erode, Tamil Nadu, India*

Abstract

Social media platforms have become a breeding ground for hostility and toxicity, with abusive language targeting women becoming a widespread issue. This paper addresses the detection of abusive content in Tamil and Malayalam social media comments using machine learning models. We experimented with GRU, LSTM, Bidirectional LSTM, CNN, FastText, and XGBoost models, evaluating their performance on a code-mixed dataset of Tamil and Malayalam comments collected from YouTube. Our findings demonstrate that the the FastText and CNN models yielded the best performance among the evaluated classifiers, achievF1 scoresores of 0.73 each. This study contributes to ongoing research on abusive text detection for under-resourced languages and highlights the need for robust, scalable solutions to combat online toxicity.

1 Introduction

The rise of social media has revolutionized how individuals communicate, share opinions, and engage with global communities. However, this unprecedented connectivity comes at the cost of an alarming increase in abusive language, particularly targeting women. Abusive language not only perpetuates gender inequality, but also has severe psychological and social consequences. Addressing this issue requires efficient tools to detect and mitigate this content effectively.

Previous works on abusive text detection have predominantly focused on English, leaving low-resource languages like Tamil and Malayalam underexplored. Moreover, the code-mixed nature of these languages further complicates the task, as traditional monolingual models fail to handle linguistic complexities inherent in such data. Building on the growing body of research on offensive language detection, this study proposes the application of machine learning models for classifying Tamil and Malayalam social media comments as abusive or non-abusive.

2 Literature Survey

The rise of social networks has required automated methods to detect and mitigate offensive content (Blair, 2003). While fostering global communication, social platforms have also become hubs for harmful language targeting individuals and groups. Advances in natural language processing (NLP) have enabled sophisticated systems to classify abusive language, even in multilingual and code-mixed contexts (Lee and Kim, 2015). However, detecting nuanced, context-dependent abuse remains challenging due to its subjective nature and linguistic variations (Obadimu, 2020).

Early studies demonstrated the effectiveness of machine learning models like Support Vector Machines (SVMs) and Naive Bayes, which relied on handcrafted features such as n-grams and TF-IDF. Deep learning models like CNNs and RNNs further improved classification by capturing contextual and sequential text patterns (T. De Smedt, 2018; Waseem and Hovy, 2016). Ribeiro et al. (M. H. Ribeiro and Jr, 2018) analyzed hateful behavior on Twitter using machine learning, while Kshirsagar et al. (P. Mishra and Shutova, 2018) highlighted the role of predictive embeddings in enhancing hate speech detection.

More recently, transformer-based models such as BERT and RoBERTa have set new benchmarks in offensive language detection by leveraging large-scale pretraining and fine-tuning (J. Mitrović and Granitzer, 2019; Fortuna and Nunes, 2019). These models effectively capture complex linguistic structures, making them ideal for tackling abusive language detection.

Despite these advancements, their application to low-resource and code-mixed languages, like Tamil and Malayalam, remains underexplored (C. Nobata,

2016). Code-mixed text presents challenges such as irregular grammar, mixed scripts, and context-switching, which existing models trained on high-resource languages struggle to address (Schmidt and Wiegand, 2018). Bridging this gap is essential for developing inclusive tools that curb online abuse across diverse linguistic communities.

3 Materials and Methods

3.1 Task Description

This study classifies Tamil and Malayalam social media comments as either abusive or non-abusive. The dataset consists of YouTube comments annotated with binary labels:

- Abusive
- Non-Abusive

3.2 Dataset

The dataset includes Tamil and Malayalam code-mixed comments from YouTube, annotated based on content. It consists of 5,000 comments, with an average sentence length of 1. Figure 1 presents sample texts.

Text	Label
உங்கள் முயற்சி வெற்றியடைய வாழ்த்துகள்!	Non-Abusive
നിന്നക് രോബിനെ കല്യാണം കഴിക്കണ്ടന്ന് ശരിക്കും ഉള്ള സത്യം.	Non-Abusive
நீ வெறும் வேலைக்கு ஒத்திகை இல்லாத நபர்!	Abusive
നവ്യയുടേ കയ്യിന് കീഴിലല്ലോ. ഇവാക്ക് അതോടെ എറിയില്ലേ	Abusive

Figure 1: Sample training texts from the dataset are shown below.

3.3 Preprocessing and Feature Extraction

Preprocessing was essential for effective classification and involved:

3.3.1 Text Cleaning

Noise such as punctuation, special characters, and emojis was removed. Emojis were converted into textual descriptions to retain sentiment (J. Salminen, 2020).

3.3.2 Tokenization

Text was split into individual tokens, allowing models to analyze semantic patterns and relationships (Gao and Huang, 2020).

3.3.3 Feature Extraction

TF-IDF vectorization assigned weights to words based on frequency, ensuring focus on informative features (H. Mubarak and Magdy, 2017; A. Vidgen, 2020). This transformation structured the data for machine learning models (Agrawal and Awekar, 2018).

3.4 Models

We evaluated various models for abusive content detection:

- **GRU**: A recurrent neural network (RNN) capturing text dependencies efficiently.
- **LSTM**: Addresses the vanishing gradient problem, effectively handling long-range dependencies.
- **Bidirectional LSTM**: Enhances context understanding by processing sequences in both directions.
- **CNN**: Extracts n-gram-like features, making it computationally efficient for classification.
- **FastText**: Embedding-based model averaging word vectors for classification.
- **XGBoost**: Gradient boosting framework leveraging decision trees for structured data classification.

4 Results and Discussion

4.1 Performance Metrics

The models were evaluated using **Accuracy, Precision, Recall, and F1-Score**, as summarized in Figure 2.

These commonly used evaluation metrics are defined as follows:

- **Accuracy**: The proportion of correctly classified texts:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- **Recall (Sensitivity)**: The proportion of correctly classified texts in a class:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

- **Precision (Positive Predictive Value)**: The proportion of correct predictions per class:

Model	Precision	Recall	F1-Score	Accuracy
GRU	0.69	0.69	0.69	69%
LSTM	0.69	0.69	0.69	69%
Bidirectional LSTM	0.26	0.50	0.34	52%
CNN	0.73	0.73	0.73	73%
FastText	0.73	0.73	0.73	73%
XGBoost	0.68	0.67	0.67	67%

Figure 2: **Performance Metrics Table.**

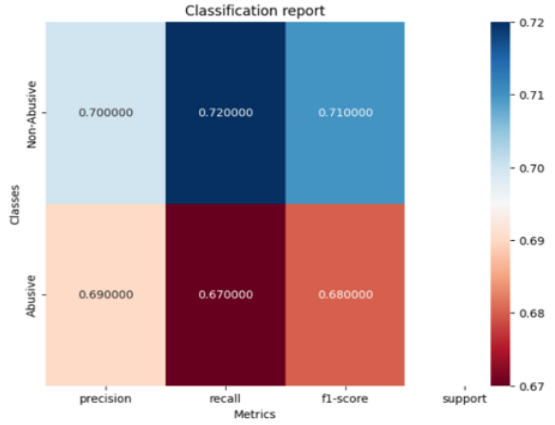
$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

- **F1-Score**: The harmonic mean of Precision and Recall:

$$\text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

4.2 Model Performance Analysis

The classification performance was analyzed using these metrics. The detailed reports for selected models are shown below (see Figures 3, 4, 5, and 6).



(a) GRU Model

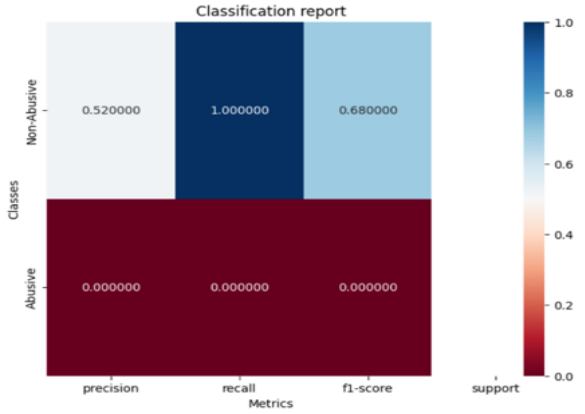
Figure 3: **GRU Model Performance**

5 Error Analysis

To better understand the challenges in detecting abusive content, we performed both qualitative and quantitative error analysis.

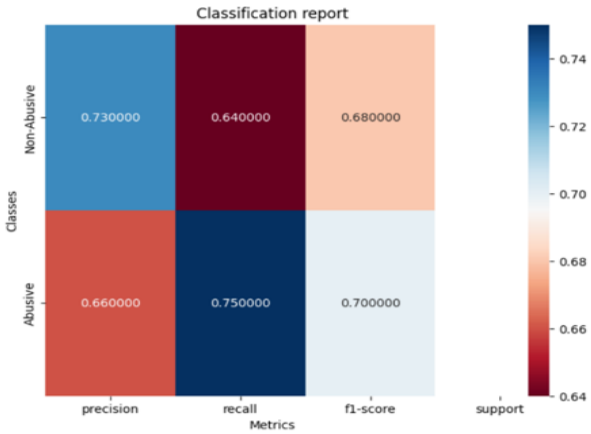
5.1 Qualitative Analysis

We manually inspected misclassified examples to identify patterns. Some key observations include:



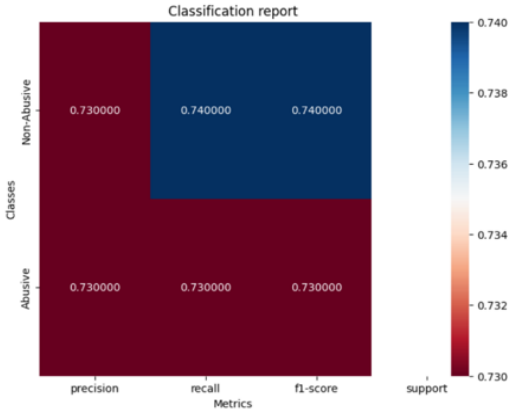
(c) Bidirectional LSTM

Figure 4: **Bidirectional LSTM Model Performance**



(b) LSTM Model

Figure 5: **LSTM Model Performance**



(d) CNN Model

Figure 6: **CNN Model Performance**

- Code-mixed comments with informal spelling variations were often misclassified.
- Sarcasm and implicit abuse were challenging for the models to detect accurately.
- Certain abusive words were misclassified due to their different contextual meanings.

5.2 Quantitative Analysis

We examined key misclassification trends:

- The CNN and FastText models performed well but misclassified some non-abusive comments as abusive.
- GRU and LSTM models struggled with long-text dependencies, leading to errors.
- Class imbalance affected the F1-score, causing a bias toward the majority class.

This analysis highlights the need for improved preprocessing techniques and context-aware abuse detection.

5.3 Hyperparameter Settings

The hyperparameters used for training the models are summarized in Table 1.

Hyperparameter	Value
Learning Rate	0.001
Batch Size	32
Dropout Rate	0.3
Number of Epochs	10
Optimizer	Adam
Loss Function	Cross-Entropy Loss

Table 1: Hyperparameter settings used for training models.

6 Conclusion

This study conducts a comparative evaluation of machine learning models for detecting abusive content in Tamil and Malayalam social media comments. The results reveal that CNN and FastText models achieved superior performance, with each attaining an F1-Score of 0.73. These findings highlight the effectiveness of these models in addressing the complexities of code-mixed and low-resource language datasets, where traditional methods often struggle. Despite this success, there remains considerable scope for improvement(Schmidt and Wiegand, 2018; Zhang and Luo, 2018). Future work will explore cutting-edge transformer-based architectures like BERT, RoBERTa, and multilingual models, which have shown significant promise in other language processing tasks. Additionally, advanced feature representation techniques, (Chakravarthi et al., 2025) such as contextual embeddings and hybrid feature extraction methods, will be investigated to enhance the models' capability to capture nuanced and context-dependent abusive language more effectively.

7 Limitations

While our approach demonstrates promising results in detecting abusive and sentiment-based text in low-resource languages, several limitations remain:

- **Data Imbalance:** The dataset contains an uneven distribution of classes, which may lead to biased predictions, especially for underrepresented labels.
- **Code-Mixed Challenges:** Handling code-mixed text remains complex due to variations in spelling, grammar, and transliteration across languages.
- **Generalization:** The trained models may not generalize well to unseen datasets or different social media platforms due to variations in language usage.
- **Computational Constraints:** Transformer-based models require significant computational resources, making deployment on low-end devices challenging.
- **Contextual Limitations:** Certain comments require deeper contextual understanding, which current models may struggle to interpret accurately.

Project Repository

The full source code for this project is available on GitHub: GitHub Repository - Deepikagowtham

References

et al. A. Vidgen. 2020. Challenges and frontiers in abusive content detection. *arXiv preprint arXiv:2010.07395*.

S. Agrawal and A. Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Proceedings of the European Conference on Information Retrieval*.

J. Blair. 2003. New breed of bullies torment their peers on the internet. *Education Week*, 22:6.

et al. C. Nobata. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the shared task on political multiclass sentiment analysis of tamil x(twitter) comments: Dravidianlangtech@naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

A. Fortuna and S. Nunes. 2019. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 52(4).

T. Gao and M. Huang. 2020. Detecting online hate speech using context-aware models. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*.

K. Darwish H. Mubarak and W. Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*.

B. Birkeneder J. Mitrović and M. Granitzer. 2019. nlpup at semeval-2019 task 6: A deep neural language model for offensive language detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

et al. J. Salminen. 2020. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. *ACM Transactions on Social Computing*.

S.-H. Lee and H.-W. Kim. 2015. Why people post benevolent and malicious comments online. *Communications of the ACM*, 58:74–79.

Y. A. Santos V. A. Almeida M. H. Ribeiro, P. H. Calais and W. Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth International AAAI Conference on Web and Social Media*.

A. M. Obadimu. 2020. *Assessing the Role of Social Media Platforms in the Propagation of Toxicity*. Ph.D. thesis, University of Arkansas at Little Rock.

H. Yannakoudakis P. Mishra and E. Shutova. 2018. Neural character-based composition models for abuse detection. *arXiv preprint arXiv:1809.00378*.

J. Schmidt and A. Wiegand. 2018. A survey on hate speech detection using natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*.

E. Kotzé L. Saoud M. Gwóźdź G. De Pauw et al. T. De Smedt, S. Jaki. 2018. Multilingual cross-domain perspectives on online hate speech. *arXiv preprint arXiv:1809.03944*.

Z. Waseem and D. Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection. In *NAACL Student Research Workshop*, pages 88–93.

M. Zhang and Y. Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *arXiv preprint arXiv:1803.03662*.

JustATalentedTeam@DravidianLangTech 2025: A Study of ML and DL approaches for Sentiment Analysis in Code-Mixed Tamil and Tulu Texts

Ponsubash Raj R, Paruvatha Priya B, Bharathi B
Department of Computer Science and Engineering
Sri Sivasubramania Nadar College of Engineering
ponsubashraj2370043@ssn.edu.in
paruvathapriya2370053@ssn.edu.in
bharathib@ssn.edu.in

Abstract

The growing prevalence of code-mixed text on social media presents unique challenges for sentiment analysis, particularly in low-resource languages like Tamil and Tulu. This paper explores sentiment classification in Tamil-English and Tulu-English code-mixed datasets using both machine learning (ML) and deep learning (DL) approaches. The ML model utilizes TF-IDF feature extraction combined with a Logistic Regression classifier, while the DL model employs FastText embeddings and a BiLSTM network enhanced with an attention mechanism. Experimental results reveal that the ML model outperforms the DL model in terms of macro F1-score for both languages. Specifically, for Tamil, the ML model achieves a macro F1-score of 0.46, surpassing the DL model's score of 0.43. For Tulu, the ML model significantly outperforms the DL model, achieving 0.60 compared to 0.48. This performance disparity is more pronounced in Tulu due to its smaller dataset size of 13,308 samples compared to Tamil's 31,122 samples, highlighting the data efficiency of ML models in low-resource settings. The study provides insights into the strengths and limitations of each approach, demonstrating that traditional ML techniques remain competitive for code-mixed sentiment analysis when data is limited. These findings contribute to ongoing research in multilingual NLP and offer practical implications for applications such as social media monitoring, customer feedback analysis, and conversational AI in Dravidian languages.

1 Introduction

With the growing prevalence of code-mixed text on social media, there is an increasing need for effective NLP tools to analyze and interpret such data. Code-mixing, the blending of words or phrases from multiple languages within the same sentence, reflects real-world multilingual communication but introduces complexities in text analysis. Sentiment analysis of code-mixed data is particularly challenging due to the lack of standardized grammar, varying transliterations, and diverse language structures.

This paper focuses on addressing these challenges in Tamil and Tulu code-mixed sentiment analysis by comparing two different methodologies: a machine learning (ML)-based approach and a deep learning (DL)-based approach. While the ML model leverages character-level n-grams and a logistic regression classifier, the DL model employs a BiLSTM architecture enhanced with attention mechanisms to capture semantic and contextual relationships in the text.

Although the performance of the DL approach is comparable to that of the ML approach in the Tamil code-mixed dataset, the difference in performance is found to be larger in the case of the Tulu code-mixed dataset, likely due to the smaller dataset. By evaluating the performance of these methodologies, this paper provides insights into the strengths and limitations of each approach, offering guidance for future work in code-mixed NLP tasks.

The paper begins with a review of related work in sentiment analysis for code-mixed text, highlighting previous research in this domain. In Section 3, the proposed methodologies, detailing the ML and DL approaches used for analysis are presented. Finally, the results are discussed, providing key insights and inferences drawn from the comparative evaluation.

2 Related Work

The study of sentiment analysis in Dravidian code-mixed text has recently gained attention, addressing the unique challenges of low-resource languages and their complex linguistic patterns. Chakravarthi et al. (2020a) focused on developing and evaluating models for Tamil-English and Tulu-English datasets, highlighting the need for effective tools to

analyze multilingual social media content. Hegde et al. (2022) introduced a valuable dataset of annotated YouTube comments and evaluated sentiment analysis using machine learning models such as logistic regression, random forest, and BERT, providing a foundation for future research in this area.

Expanding on these efforts, Hegde et al. (2023) presented findings from a shared task on sentiment analysis in code-mixed texts. The growing interest in sentiment analysis for Dravidian code-mixed languages is further reflected in initiatives aimed at benchmarking models for multilingual datasets. Kumar et al. (2024) discuss these initiatives, highlighting advancements in sentiment analysis for social media content.

To enhance sentiment classification in code-mixed texts, Puranik et al. (2021) employed pre-trained models like ULMFiT and multilingual BERT, fine-tuning them on the code-mixed dataset, its transliteration (TRAI), an English translation (TRAA) of the transliterated data, and a combination of all three. Similarly, Balaji et al. (2020) analyzed different feature extraction techniques, including count vectorization, LSTM, and BERT embeddings, comparing their effectiveness across various machine learning models. While sentiment analysis in high-resource languages like English has been extensively studied, research on Dravidian code-mixed languages, particularly Tamil and Tulu, remains in its early stages. Ponnusamy et al. (2023) emphasize the need for robust models capable of handling informal grammar and mixed scripts, underscoring the importance of tailored approaches for low-resource languages.

Early-stage research on sentiment analysis in Dravidian languages has paved the way for further exploration of text classification and emotion detection. Rachana et al. (2023) discuss the increasing interest in analyzing code-mixed text, encouraging the development of more inclusive and diverse language processing tools. Likewise, Chakravarthi et al. (2020b) focus on sentiment analysis in Tamil-English and Malayalam-English code-mixed texts, addressing the complexities introduced by mixed scripts and informal syntax in social media data.

Various machine learning approaches have been employed to enhance sentiment classification in multilingual online content. Kanta and Sidorov (2023) examined sentiment detection in Tamil-English and Tulu-English code-mixed social media text, using machine learning techniques to improve emotion recognition in such contexts. Similarly, Bharathi and Samyuktha (2021) and Varsha et al. (2022) explored machine learning-based approaches for sentiment analysis, with the latter also incorporating transformer models to refine classification performance.

The proposed models in the paper and their predictions are submitted for the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025 Durairaj et al. (2025). Given the frequent blending of Dravidian languages with English on social media platforms, sentiment analysis remains a challenging task. Recent studies have leveraged both traditional machine learning and deep learning methods to refine emotion classification in multilingual contexts, improving sentiment detection in mixed-language user-generated content. As research in this field progresses, the development of comprehensive datasets and advanced modeling techniques will be crucial in enhancing sentiment analysis for low-resource languages.

3 Proposed Methodology

This section describes the methodology followed for sentiment classification in code-mixed Tamil and Tulu texts using both Machine Learning (ML) and Deep Learning (DL) approaches. The dataset split for training and testing is shown in Table 1.

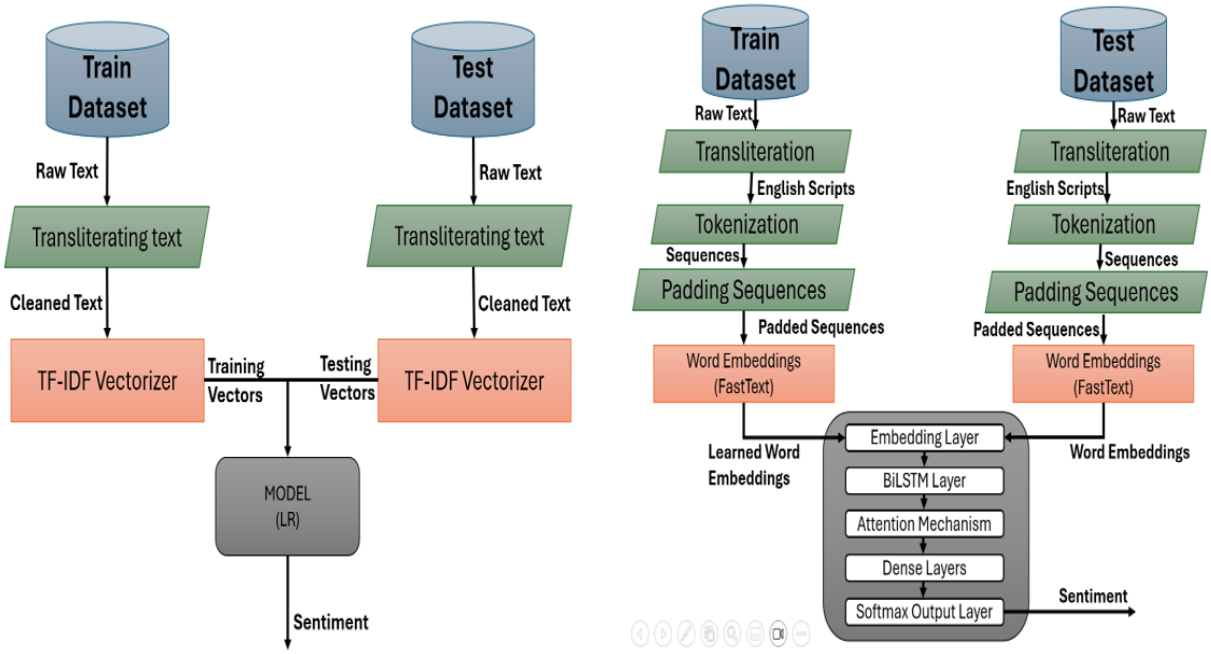
3.1 Preprocessing

Since the dataset contains code-mixed Tamil-English and Tulu-English text, transliteration is performed to convert all text into the English script for uniform processing. Following this, tokenization is applied to split text into individual words or subwords, enabling better feature extraction.

The text is then lowercased to maintain consistency. Additionally, punctuation marks and special characters are eliminated to reduce noise, and normalization techniques are used to handle repeated characters. Once preprocessed, the cleaned text is used as input for both ML and DL models, with different feature extraction techniques applied in each approach.

3.2 Machine Learning Approach

In the ML approach, numerical feature representations of text are generated using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer, which captures the importance of words within the dataset. The TF-IDF vectorization is



(a) Machine Learning Approach

(b) Deep Learning Approach

Figure 1: Architecture diagram of the proposed methodology using (a) machine learning and (b) deep learning approaches.

Table 1: Data Distribution

Language	Training Set	Validation Set
Tamil	31122	3843
Tulu	13308	1643

performed separately on the training and testing datasets, using a character-level analyzer with an n-gram range of (1,4) to effectively capture word-level and subword-level patterns which is vital for code-mixed texts.

Once the text is converted into feature vectors, a Logistic Regression (LR) classifier is trained using these numerical representations. The model is trained with the 'liblinear' solver. The trained model is then used to predict sentiment labels for the test dataset. The performance is evaluated using accuracy and macro F1-score, with the given validation dataset. This approach relies on statistical feature extraction and performs effectively with relatively smaller datasets, making it well-suited for sentiment classification in code-mixed text. The overall architecture of the ML approach is shown in Figure 1a.

3.3 Deep Learning Approach

The DL approach leverages word embeddings and sequential modeling to better capture the contextual relationships within the text. The cleaned text is first tokenized and converted into sequences, which are then mapped to word embeddings using FastText, trained on the dataset to generate dense vector representations for words, including out-of-vocabulary words. The sequences are padded to 100 tokens. The embeddings are initialized with a dimension of 300 and a window size of 5.

These embeddings serve as input to a Bidirectional LSTM (BiLSTM) model with an attention mechanism, allowing the model to capture both forward and backward dependencies in the text. The BiLSTM layer consists of 128 hidden units and includes a dropout rate of 0.3 to prevent overfitting. The attention layer refines the model's focus on the most informative words, followed by fully connected dense layers with ReLU activation and L2 regularization. The final softmax layer classifies the text into sentiment categories. The model is trained using sparse categorical cross-entropy loss and optimized with the Adam optimizer, set with an initial learning rate of 0.001 with decay. The overall architecture of the Deep Learning approach is shown in Figure 1b.

Table 2: Performance analysis of the proposed system using validation data

Language	Model	Macro F1 Score	Accuracy
Tamil	Logistic Regression	0.46	0.54
Tamil	BiLSTM with Attention	0.43	0.58
Tulu	Logistic Regression	0.60	0.69
Tulu	BiLSTM with Attention	0.48	0.65

While the DL approach can capture deeper semantic relationships, its performance is dependent on the availability of large labeled datasets. In this study, the ML approach outperforms the DL model, particularly for the Tulu dataset, indicating that traditional feature-based methods may still be effective in low-resource settings where deep learning models struggle with limited training data. Additionally, the smaller dataset size may have led to suboptimal generalization in the deep learning model.

4 Results

The experimental results from Table 2 reveal that the ML model (TF-IDF + Logistic Regression) outperforms the DL model (FastText + BiLSTM + Attention) for both Tamil and Tulu datasets in terms of macro F1 scores. For Tamil, the ML model achieved a score of **0.46** compared to **0.43** for the DL model, while for Tulu, the ML model performed significantly better with a score of **0.60** compared to **0.48** for the DL model. The disparity in performance is more pronounced for Tulu, likely due to its smaller dataset size (13,308 samples) compared to Tamil (31,122 samples). These results highlight the effectiveness of ML models in handling limited data, whereas the DL model struggled due to its reliance on larger datasets for effective representation learning. The code used for these experiments is available on GitHub.[1]

5 Conclusions

The findings demonstrate that ML models are better suited for sentiment analysis of code-mixed texts, particularly in low-resource settings, as they effectively leverage n-gram-based features without requiring extensive labeled data. DL models, while theoretically capable of capturing richer contextual and semantic relationships, underperform with limited data availability. This is because deep learning models require a large amount of training data to learn meaningful representations and avoid overfitting, whereas smaller datasets fail to provide sufficient examples for learning complex patterns. This study emphasizes the need for larger and more diverse datasets to fully realize the potential of DL models for code-mixed text analysis and suggests exploring transfer learning or pre-trained multilingual models to improve performance in low-resource scenarios. Overall, ML models remain a practical and reliable approach for code-mixed sentiment analysis, especially for underrepresented languages like Tulu.

6 Limitations

This study is limited by the dataset size, particularly for Tulu ($13,308$ samples), which affects model generalizability. It focuses only on Tamil-English and Tulu-English code-mixed texts, limiting applicability to other Dravidian languages. The ML model uses TF-IDF, which lacks contextual understanding, while the DL model (FastText + BiLSTM + Attention) is not fine-tuned on Dravidian corpora. Transformer-based models like mBERT and XLM-R are not explored, and transliteration variations are not explicitly handled. Addressing these limitations in future work could improve sentiment classification in Dravidian code-mixed texts.

[1] https://github.com/JustATalentedGuy/JustATalentedTeam_NAACL

References

Nitin Nikamanth Appiah Balaji, B Bharathi, and J Bhuvana. 2020. Ssncse_nlp@ dravidian-codemix-fire2020: Sentiment analysis for dravidian languages in code-mixed text. In *FIRE (Working Notes)*, pages 554–559.

B Bharathi and GU Samyuktha. 2021. Machine learning based approach for sentiment analysis on multilingual code mixing text. In *FIRE (Working Notes)*, pages 1038–1043.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P McCrae. 2020a. Corpus creation for sentiment analysis in code-mixed tamil-english text. *arXiv preprint arXiv:2006.00206*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020b. Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 21–24.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, SK Lavanya, Durairaj Thenmozhi, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71.

Selam Kanta and Grigori Sidorov. 2023. Selam@ dravidianlangtech: Sentiment analysis of code-mixed dravidian texts using svm classification. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 176–179.

Lavanya Sambath Kumar, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 2024. Overview of second shared task on sentiment analysis in code-mixed tamil and tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 62–70.

Kishore Kumar Ponnusamy, Charmathi Rajkumar, Prasanna Kumar Kumaresan, Elizabeth Sherly, and Ruba Priyadharshini. 2023. Vel@ dravidianlangtech: Sentiment analysis of tamil and tulu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 211–216.

Karthik Puranik et al. 2021. Iiitt@ dravidian-codemix-fire2021: Transliterate or translate? sentiment analysis of code-mixed text in dravidian languages. *arXiv preprint arXiv:2111.07906*.

K Rachana, M Prajnashree, Asha Hegde, and HL Shashirekha. 2023. Mucs@ dravidian-langtech2023: Sentiment analysis in code-mixed tamil and tulu texts using fasttext. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 258–265.

Josephine Varsha, B Bharathi, and A Meenakshi. 2022. Sentiment analysis and homophobia detection of youtube comments in code-mixed dravidian languages using machine learning and transformer models. In *FIRE (Working Notes)*, pages 124–137.

KEC_TECH_TITANS@DravidianLangTech 2025:Sentiment Analysis for Low-Resource Languages: Insights from Tamil and Tulu using Deep Learning and Machine Learning Models

Malliga Subramanian[1], Kogilavani S V[1], Dharshini S[1], Deepiga P[1], Praveenkumar C[1], Ananthakumar S[1]

[1]*Kongu Engineering College, Erode, Tamil Nadu, India*

Abstract

Sentiment analysis in Dravidian languages like Tamil and Tulu presents significant challenges due to their linguistic diversity and limited resources for natural language processing (NLP). This study explores sentiment classification for Tamil and Tulu, focusing on the complexities of handling both languages, which differ in script, grammar, and vocabulary. We employ a variety of machine learning and deep learning techniques, including traditional models like Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), as well as advanced transformer-based models like BERT and multilingual BERT (mBERT). A key focus of this research is to evaluate the performance of these models on sentiment analysis tasks, considering metrics such as accuracy, precision, recall, and F1-score. The results show that transformer-based models, particularly mBERT, significantly outperform traditional machine learning models in both Tamil and Tulu sentiment classification. This study underscores the need for further research to overcome challenges like language-specific nuances, dataset imbalance, and data augmentation techniques for improved sentiment analysis in under-resourced languages like Tamil and Tulu.

1 Introduction

Sentiment analysis in Tamil and Tulu is challenging due to their linguistic diversity and the lack of annotated datasets. Tamil is spoken in India and Sri Lanka, while Tulu is mainly used in coastal Karnataka and Kerala. Their unique linguistic features make sentiment analysis crucial for opinion mining, social media monitoring, and customer feedback analysis. The challenge increases with code-mixing, where users frequently switch between Tamil, Tulu, and English, especially on social media, making it harder to train and evaluate models effectively. Recent advancements in machine learning and deep learning have improved sentiment analysis, with transformer-based models like BERT and mBERT achieving state-of-the-art results. However, their application in Tamil and Tulu remains underexplored (Durairaj et al., 2025). This study evaluates traditional models such as Logistic Regression, SVM, and KNN, alongside advanced models like BERT and mBERT, for sentiment classification in Tamil and Tulu. We analyze how well they handle code-mixing and data scarcity, offering insights to improve sentiment analysis in low-resource languages and develop better NLP tools for Tamil and Tulu. Additionally, this research aims to bridge the gap in sentiment analysis for Dravidian languages by exploring more effective machine learning approaches.

2 Literature Survey

Sentiment analysis is a vital area of research in natural language processing (NLP), with significant applications in opinion mining, social media monitoring, and customer feedback analysis. While considerable progress has been made in sentiment classification for high-resource languages like English, research for low-resource languages, particularly Dravidian languages such as Tamil and Tulu, remains underexplored. The complexity of sentiment analysis in these languages arises from their diverse linguistic structures, code-switching, and cultural nuances, which make it a challenging task for machine learning and deep learning models. Additionally, the lack of large, annotated datasets in Tamil and Tulu further complicates the development of robust models for sentiment classification.

Thenmozhi et al. (2025) provide an extensive overview of sentiment analysis for Tamil and Tulu, detailing the challenges and methods used in previous research (Durairaj et al., 2025). Their study emphasizes the significance of transformer-based models, particularly multilingual BERT (mBERT),

in improving sentiment classification performance. The work also highlights the difficulties associated with code-mixing, data scarcity, and the necessity of annotated datasets for effective sentiment analysis in low-resource languages.

2.1 Sentiment Analysis in Tamil

Sentiment analysis in Tamil has evolved significantly over the years. Early studies relied on traditional machine learning models like Support Vector Machines (SVM) and Naive Bayes (NB), using features such as n-grams and sentiment lexicons. (Prabhu and Sundararajan, 2014) achieved moderate success with accuracy rates of 70-80%, but the use of manually created resources and limited ability to capture context restricted model performance.

In recent years, deep learning approaches have been adopted, offering improvements in understanding contextual meaning. (Babu and Ranjan, 2020) used Convolutional Neural Networks (CNN) for sentiment analysis in Tamil, yielding better results compared to traditional methods. More recently, (Ranjan and Babu, 2021) applied BERT-based models, which outperformed previous methods due to their ability to learn contextual embeddings, marking a significant advancement in Tamil sentiment analysis.

2.2 Sentiment Analysis in Tulu

Sentiment analysis in Tulu, a low-resource Dravidian language, has been underexplored due to limited linguistic resources and annotated corpora. Early attempts at Tulu sentiment classification were largely reliant on traditional machine learning models, such as Support Vector Machines (SVM) and Naive Bayes (NB), often utilizing handcrafted features like n-grams and lexicon-based approaches. These methods faced challenges due to the lack of large-scale labeled datasets, and their performance was constrained by the language's unique syntactic structure and informal usage in social media and online platforms. Additionally, rule-based approaches and domain-specific lexicons were employed, but they struggled to capture the complexities and nuances of Tulu sentiment expressions.

In recent years, there has been a shift towards leveraging deep learning models, particularly transformer-based architectures like multilingual BERT (mBERT), which are pre-trained on large multilingual corpora and can be fine-tuned for low-resource languages such as Tulu. These models, with their capacity to learn contextual embeddings and capture long-range dependencies, have demonstrated superior performance in sentiment analysis for other Dravidian languages, such as Tamil and Malayalam. Transfer learning, where models trained on high-resource languages are adapted to Tulu, is emerging as a promising strategy to overcome data limitations. However, the scarcity of annotated Tulu datasets remains a significant challenge, underscoring the need for further research and the development of comprehensive labeled corpora to improve the robustness and accuracy of sentiment classification systems for Tulu.

2.3 Challenges in Sentiment Analysis for Dravidian Languages

Several challenges hinder the progress of sentiment analysis in Dravidian languages. One of the key difficulties is the lack of large, annotated datasets, which are essential for training robust models. Dravidian languages have diverse morphological structures, dialects, and variations in colloquial language, which further complicates sentiment detection.

Another significant challenge is code-switching, where speakers alternate between Dravidian languages and languages like English, especially in digital communication. (Subashini et al., 2022) found that sentiment analysis in code-mixed data leads to a decline in model performance, as models trained on single-language datasets struggle to interpret the multilingual input effectively.

2.4 Recent Advances and Transformer Models

Recent advancements in NLP, particularly with transformer-based models such as BERT and its multilingual variant mBERT, have shown promising results in sentiment analysis tasks for low-resource languages. (Vaswani et al., 2017) introduced the Transformer architecture, which revolutionized NLP tasks by providing a method to capture long-range dependencies in text without relying on sequential processing, as was the case with RNNs and LSTMs.

In the context of Dravidian languages, (Ghosal et al., 2020) demonstrated the utility of mBERT in multilingual sentiment analysis tasks, where it outperformed traditional machine learning approaches by leveraging cross-lingual transfer learning. This has opened new avenues for handling sentiment analysis in under-resourced languages like Tamil

and Tulu, as mBERT is pre-trained on a large corpus of multilingual data, providing a foundation to fine-tune models for specific languages or tasks.

3 Materials and Methods

3.1 Dataset Description

The dataset used in this study consists of Tamil-Tulu code-mixed text collected from social media platforms such as Twitter, Facebook, and regional online forums. It contains 8,000 training samples and 2,000 validation samples. The sentiment class distribution in the Tamil dataset is as follows: 58.30% Positive, 13.34% Negative, and 11.77% Mixed Feelings, with an additional 16.59% categorized as "Unknown State" that required filtering. In the validation set, the distribution remains similar, with 59.12% Positive, 12.49% Negative, and 12.28% Mixed Feelings, alongside 16.11% Unknown State.

For the Tulu dataset, a significant proportion (33.08% in training and 33.07% in validation) consists of "Not Tulu" samples, which were excluded from sentiment classification. Among the actual sentiment labels, 28.34% were Positive, 23.87% Neutral, 8.38% Mixed, and 6.34% Negative in the training set, while the validation set showed 28.62% Positive, 22.41% Neutral, 8.71% Mixed, and 7.19% Negative. The dataset also exhibited class imbalance, particularly in the Mixed category, which was addressed using oversampling techniques like SMOTE and random resampling.

3.2 Pre-processing and Feature Extraction

Preprocessing and feature extraction for sentiment analysis in Tamil and Tulu involve several key steps to handle the unique linguistic challenges of these languages. Text normalization standardizes the text by lowercasing, removing extra spaces, and handling informal contractions, while tokenization splits the text into words or subwords, especially in code-mixed contexts.

Stopwords and noise, such as emojis, special characters, and URLs, are removed to reduce irrelevant information. Lemmatization or stemming is applied to reduce words to their base forms, and code-switching between Tamil, Tulu, and English is handled using language identification and transliteration techniques. Feature extraction includes techniques like Bag-of-Words (BoW) and TF-IDF to capture word frequencies, as well as more advanced approaches like word embeddings (Word2Vec, FastText) and transformer-based models (BERT, mBERT) to capture semantic meaning and contextual nuances. Additionally, sentiment lexicons specific to Tamil and Tulu can be integrated to enhance the detection of sentiment-bearing words. These preprocessing and feature extraction steps enable effective sentiment classification by preparing the data for machine learning and deep learning models.

3.3 Proposed Classifiers

To classify sentiment in Tamil and Tulu, we used both traditional machine learning and deep learning models. SVM and KNN were chosen for their efficiency with small datasets and clear decision boundaries, while Decision Trees helped explore rule-based classification but struggled with complex language structures and code-mixing.

Deep learning models like CNNs and RNNs captured patterns and sequential dependencies, making them better suited for morphologically rich languages. However, transformer-based models like BERT and mBERT performed best due to their strong contextual understanding and multilingual capabilities. Given the frequent code-mixing in Tamil-Tulu, mBERT excelled due to its pre-training on diverse languages. This combination of models helped balance context, sequence learning, and feature-based classification to improve sentiment analysis accuracy.

4 Results and Discussion

The sentiment analysis results for Tamil and Tulu indicate that transformer-based models like BERT and mBERT delivered the highest accuracy due to their ability to capture complex contextual information in code-mixed text. CNNs effectively detected local patterns, while RNNs and GRUs excelled in handling sequential dependencies across sentences. MLPs provided reasonable performance, while Random Forests offered robustness but slightly lower accuracy. Logistic Regression, though efficient for binary classification, performed less effectively than deep learning models. Combining these approaches enhanced overall accuracy.

4.1 Performance Metrics

The performance metrics for sentiment analysis in Tamil and Tulu include accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC curve. Accuracy measures the overall proportion of correct classifications, while precision evaluates the

ability to avoid false positives, and recall assesses the model's capability to capture all relevant sentiment instances. The F1-score balances precision and recall, providing a comprehensive measure. The confusion matrix breaks down classification errors, and the AUC-ROC curve indicates the model's ability to distinguish between different sentiment classes. These metrics offer a thorough evaluation of the model's effectiveness, especially in the context of code-mixed text and language-specific nuances.

Table 1: Performance Table of our Models for Sentiment Analysis in Tamil and Tulu

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
KNN	56	64	63	64
SVM	77	83	80	83
DT	73	68	65	66
BERT	85	80	78	82
RNN	70	65	60	63

4.2 Limitations

While our approach improves sentiment classification for Tamil and Tulu, several challenges remain.

First, dataset imbalance affects performance, especially in the Mixed Feelings category. Although oversampling helped, advanced data augmentation techniques could further improve balance. Second, high computational costs of deep learning models like BERT and mBERT make real-time sentiment analysis challenging. Using lighter transformer models or compression techniques could address this issue.

Another challenge is code-mixed text, where Tamil, Tulu, and English are mixed in a single sentence. While mBERT performed well, it was not trained specifically for Tamil-Tulu mixing, causing occasional errors. Future work could involve fine-tuning transformers on Tamil-Tulu datasets. Finally, limited annotated data affects accuracy. Expanding the dataset with more labeled examples and domain-specific lexicons would improve performance.

5 Conclusion and Future Work

In conclusion, this study highlighted the effectiveness of machine learning and deep learning models, particularly transformer-based models like BERT and mBERT, for sentiment analysis in Tamil and Tulu. These models outperformed traditional methods by capturing contextual nuances in code-mixed text. While CNNs, RNNs, and GRUs showed strong performance in identifying patterns and sequential dependencies, simpler models like Logistic Regression were less effective. Future work can focus on expanding the dataset, addressing class imbalance, integrating domain-specific lexicons, and fine-tuning multilingual models like mBERT to further improve sentiment analysis accuracy and model generalization for Tamil and Tulu.

Project Repository

The full source code for this project is available on GitHub: GitHub Repository Deepikagowtham

References

D. Babu and S. Ranjan. 2020. Sentiment Analysis of Tamil Text Using Convolutional Neural Networks. In *Proceedings of the 5th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2020)*.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

D. Ghosal, M. Thakur, and A. Patel. 2020. Multilingual BERT for Sentiment Analysis in Low-Resource Dravidian Languages. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*.

A. Prabhu and V. Sundararajan. 2014. Sentiment Classification for Tamil Language Using Machine Learning Algorithms. *International Journal of Computer Applications*, 97(5):28–32.

S. Ranjan and D. Babu. 2021. Exploring BERT for Sentiment Analysis in Tamil: A Deep Learning Approach. In *Proceedings of the 6th International Conference on Artificial Intelligence and Data Science (AIDS 2021)*.

R. Subashini, P. Kumar, and L. Devi. 2022. Challenges in Sentiment Analysis of Code-Mixed Dravidian Languages. *International Journal of Computational Linguistics*, 14(2):102–118.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 5998–6008.

Code_Conquerors@DravidianLangTech 2025: Multimodal Misogyny Detection in Dravidian Languages Using Vision Transformer and BERT

Pathange Omkareshwara Rao, Harish Vijay V, Ippatapu Venkata Srichandra, Neethu Mohan, Sachin Kumar S
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.u4aie22039@cb.students.amrita.edu, harishvijay0204@gmail.com,
ippatapuvenkatasrichandra@gmail.com, s_sachinkumar@cb.amrita.edu

Abstract

This research focuses on misogyny detection in Dravidian languages using multimodal techniques. It leverages advanced machine learning models, including Vision Transformers (ViT) for image analysis and BERT-based transformers for text processing. The study highlights the challenges of working with regional datasets and addresses these with innovative preprocessing and model training strategies. The evaluation reveals significant improvements in detection accuracy, showcasing the potential of multimodal approaches in combating online abuse in underrepresented languages.

Keywords: Misogyny Detection, Dravidian Languages, Tamil, Malayalam, Multimodal Techniques, Vision Transformers (ViT), BERT-Based Models, Regional Datasets, Machine Learning, Online Abuse.

1 Introduction

The study explores the underrepresented area of misogyny detection in Dravidian languages, particularly Tamil and Malayalam. Online misogyny is a growing social issue, with an increasing volume of content targeting women on digital platforms.This work was submitted to the Misogyny Meme Detection - DravidianLangTech@NAACL 2025 competition, organized by DravidianLangTech, where it achieved a **rank of 13 in Malayalam and 15 in Tamil languages**, demonstrating the effectiveness of the proposed multimodal approach in detecting misogyny in Dravidian languages. Existing methods often fail to address the nuances of regional languages and multimodal content. This research builds upon prior works that used unimodal or traditional models and introduces a robust multimodal approach. By fusing features from visual and textual data, the proposed method seeks to improve detection accuracy. The paper presents a detailed examination of the challenges, including data inconsistencies and the need for advanced preprocessing techniques, while providing solutions to these issues.Code available at: GitHub repository.

2 Literature Review

Another methodology was proposed by Gu et al. (2022a), where they introduced a multi-modal and multi-task Variational Autoencoder (VAE) on the same dataset. Their method achieved F1-scores of 0.72 for Task A and 0.634 for Task B. Despite these encouraging results, the study lacked statistical analysis of the model's outcomes, which limited its scope for deeper insights.The work by Singh et al. (2023) investigated the MAMI dataset using various state-of-the-art models. They combined a pretrained BERT model with ViT, achieving an F1-score of 0.874. However, the methodology faced limitations such as the lack of interpretability and the extended time required for training and deployment.

The authors Mahesh et al. (2024) focused on the LT-EDI @EACL 2024 dataset, applying mBERT+ResNet-50 and MuRIL+ResNet-50 models. These approaches achieved F1-scores of 0.73 and 0.87 for Tamil and Malayalam datasets, respectively.

Another methodology proposed by Gu et al. (2022b) combined joint image and text classification, resulting in a macro F1-score of 0.665. However, their work did not explore varying threshold values or class weights, and they struggled to develop an effective model that leveraged image data effectively.The authors Shanmugavadivel et al. (2023) explored abusive comment detection data and applied various machine learning, deep learn-

ing, and transformer-based approaches. Among their methods, the Random Forest model achieved a macro F1-score of 0.42. A key limitation of their work was the lack of experimentation with different pretrained BERT models.

Another methodology was proposed by Shaun et al. (2024), who used a multinomial Naive Bayes approach for textual data and a ResNet-50 model for pictorial data. Their approach on Tamil and Malayalam datasets achieved an F1-score of 0.82. Similarly, Koutlis et al. (2023) introduced a new deep-learning-based architecture called Memefier, which was tested on datasets such as Facebook Hateful Memes, Memotions 7k, and MultiOFF.

Finally, another methodology was developed by Boinepelli et al. (2020), who applied various machine learning and deep learning methods to the SemEval-2020 dataset. Their CNN-LSTM model achieved an F1-score of 59.04

3 Data Description

The dataset used for the misogyny detection task focuses on two Dravidian languages: Tamil Ponnusamy et al. (2024) Chakravarthi et al. (2024) and Malayalam Ponnusamy et al. (2024) Chakravarthi et al. (2024). It is divided into development and training datasets, each containing images and corresponding metadata in Excel sheets. The metadata includes three columns: `image_id`, `label`, and `transcription`. Here, label '1' indicates misogynistic content, while label '0' represents non-misogynistic content. However, discrepancies were observed in the original dataset between the number of images in the folders and the entries in the Excel sheets. To address this, consolidated datasets were created by aligning the images with their corresponding metadata. Below, we describe both the original and consolidated datasets.

3.1 Original Development Dataset

Folder	Images	Metadata	Labels
Malayalam	160	160	0: 97, 1: 63
Tamil	282	284	0: 210, 1: 74

Table 1: Details of the original development dataset.

Kumar et al. (2017) The table-1 contained separate folders for Tamil and Malayalam memes. Each folder consisted of images and metadata entries, with slight inconsistencies in the Tamil dataset between the image folder count and Excel sheet count.

3.2 Original Training Dataset

Folder	Images	Metadata	Labels
Malayalam	639	640	0: 381, 1: 259
Tamil	1135	1136	0: 851, 1: 285

Table 2: Details of the original training dataset.

The table-2 showed a larger discrepancy, particularly in the Tamil subset, with 1135 images in the folder but 1136 entries in the metadata. Despite this, the dataset provided a robust training foundation.

3.3 Consolidated Dataset

Folder	Images	Metadata	Labels
Malayalam	460	460	0: 252, 1: 168
Tamil	787	787	0: 591, 1: 196

Table 3: Details of the consolidated dataset.

The inconsistencies in the original dataset were resolved by aligning images with their respective metadata. The table-3 forms the primary input for model training and evaluation (Chakravarthi et al., 2025). The figure-1 illustrates the overall architecture of the multimodal approach used for misogyny detection in this research. The model takes in both image and text data, which undergo respective preprocessing steps such as resizing, normalization, tokenization, and padding. The preprocessed data is then fed into specialized feature extractors - Vision Transformer (ViT) for images and BERT/RoBERTa for text. The extracted features are then fused and passed through a classification layer to predict whether the input is misogynistic or non-misogynistic. The details of the individual components and training setup are described in the Methodology section.

4 Proposed Methodology

The methodology integrates preprocessing, training setup, and architecture to detect misogynistic content in memes effectively. The figure-1 shows the proposed methodology.

4.1 Architecture

Our multimodal approach processes image and text data simultaneously. The Vision Transformer Remya et al. (2024) serves as the image encoder, processing image patches as sequences for global context extraction. For text encoding, BERT (Malay-

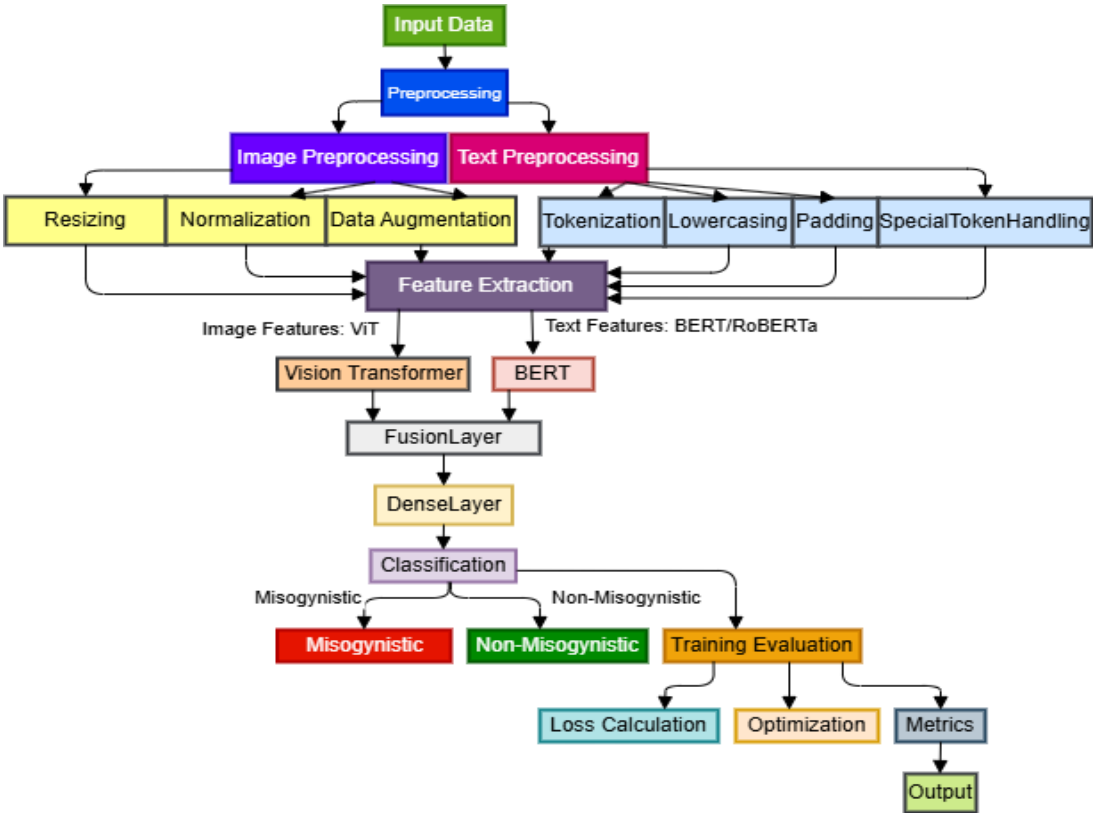


Figure 1: Overview of the Multimodal Misogyny Detection Model.

Image Model	Text Model	F1 Score (Malayalam)	F1 Score (Tamil)	Data Type
ViT	BERT-base-uncased	**0.882251**	**0.822728**	Validation
Swin Transformer	XLM-RoBERTa	0.881944	0.808841	Validation
ResNet50	RoBERTa-base	0.822145	0.797273	Validation
EfficientNet-B3	XLM-RoBERTa	0.779364	0.789389	Validation
EfficientNet-B0	DeBERTa-base	0.757143	0.779304	Validation

Table 4: Performance comparison of top models for Malayalam and Tamil datasets.

alamAbeera et al. (2023)) or RoBERTa (Tamil) extracts semantic features from captions. The features from both encoders are concatenated through a fully connected layer, forming a unified representation that feeds into a classification layer for binary prediction (misogynistic/non-misogynistic).

4.2 Preprocessing

4.2.1 Image Preprocessing

Images are resized to 224x224 pixels, normalized to [0,1], and augmented using random rotations, flips, color jittering, and cropping to prevent overfitting.

4.2.2 Text Preprocessing

Text undergoes tokenization with model-specific tokenizers (BERT, RoBERTa), uniform padding, lowercasing, and special token addition ([CLS], [SEP]).

4.3 Training Setup

Training utilizes CUDA-enabled GPUs with 15 epochs, learning rates of 2×10^{-5} (text) and 3×10^{-4} (image), binary cross-entropy loss, Adam optimizer, and batch size of 32.

4.4 Models

For Malayalam, **ViT + BERT-base-uncased** achieved the highest F1 score of **0.882251**, while for Tamil, **ViT + RoBERTa-base** scored **0.822728**.

5 Results and Discussion

5.1 Experimental Setup

The model hyperparameters (table-5) were selected through experimentation to balance performance and efficiency. A learning rate of 2e-5 was chosen for the AdamW optimizer, ensuring stable convergence for BERT-based text encoders, while a higher rate of 3e-4 was used for image encoders

Hyperparameter	Value
Learning Rate	2e-5
Batch Size	16
Dropout Rate (Fusion)	0.5
Dropout Rate (Classifier)	0.3
Fusion Dimension	512
Weight Decay	AdamW
Number of Epochs	15
Scheduler Factor	0.5
Scheduler Patience	2

Table 5: Model Hyperparameters.

(ViT/ResNet-50) to accommodate their larger parameter space. The batch size of 16 was tailored to GPU memory constraints, maintaining gradient stability. The fusion module combines text and visual embeddings into a 512-dimensional space, with a dropout rate of 0.5 to prevent overfitting. The classifier applies an additional dropout rate of 0.3 to enhance generalization. These dropout rates were empirically determined to balance regularization and feature retention. Training was conducted for 15 epochs, with a ReduceLROnPlateau scheduler reducing the learning rate by half if the validation F1-score stagnated for two epochs. This adaptive approach ensured robust convergence. Ablation studies confirmed that alternative configurations (e.g., higher dropout or larger batch sizes) resulted in lower validation F1-scores, validating the chosen hyperparameters.

5.2 Results

Data	Malayalam	Tamil
Validation	0.9115	0.8079
Test	0.75649	0.66142

Table 6: Model Performance Across Languages. *Note: All values represent F1 scores.*

The table-6 shows our evaluation of misogyny detection models, we focused on the F1 score as the primary performance metric. Our experimentation involved multimodal combinations, including ResNet-50 + BERT-base uncased and Vision Transformer (ViT) + BERT-base uncased, to assess their effectiveness in detecting misogyny in text and images.For the Tamil language Selvan et al. (2015), the macro F1 scores obtained were: the ResNet-50 + BERT-base uncased combination achieved a validation F1 score of 0.8079 and a test F1 score of 0.66142. In the case of Malayalam, the Vision Transformer (ViT) + BERT-base uncased combination achieved a validation F1 score of 0.9115 and a test F1 score of 0.75649.

5.3 Performance Comparison:

The performance of the top models for Malayalam and Tamil datasets is summarized in Table-4. The ViT + BERT-base-uncased model achieved the highest validation F1 score of 0.882 for Malayalam and 0.823 for Tamil. However, the test performance (shown in Table-6) revealed a drop in F1 scores, indicating potential overfitting or a domain gap between the training and test datasets.

6 Discussion

From Table-4 we could analyse that Tamil, ResNet-50 + BERT-base uncased performed best, but the drop in test performance suggests overfitting or a domain gap. For Malayalam, ViT + BERT-base uncased showed consistent performance, highlighting ViT's ability to capture visual details and BERT's multilingual text processing. The results emphasize the importance of selecting appropriate models for multimodal tasks and reaffirm the potential of multimodal approaches in addressing misogyny detection in regional languages. Upon examining misclassified cases, we observed that the model struggled with ambiguous memes where the text and image conveyed conflicting messages. For instance, memes with sarcastic or culturally specific humor were often misclassified. This highlights the need for better contextual understanding and cultural nuance in future models.

7 Conclusion

This study enhances misogyny detection in Dravidian languages using a multimodal approach. ResNet-50 + BERT-base uncased achieved F1 scores of 0.8079 (validation) and 0.66142 (test) for Tamil, while ViT + BERT-base uncased scored 0.9115 and 0.75649 for Malayalam. ViT captured visual patterns effectively, and BERT-base uncased handled multilingual text well. The performance gap highlights generalization challenges. Future work includes leveraging mBERT/IndicBERT, reducing domain gaps, and refining multimodal fusion to foster inclusive digital spaces.

8 Limitations

Our proposed methodology faced limitations in handling different domains, such as Tamil and Malayalam. mBERT or IndicBERT could have provided better contextual understanding and improved generalization.

References

Yimeng Gu, Ivan Castro, and Gareth Tyson. Mmvae at semeval-2022 task 5: A multi-modal multi-task vae on misogynous meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 700–710, 2022a.

S. Singh, A. Haridasan, and R. Mooney. "female astronaut: Because sandwiches won't make themselves up there": Towards multimodal misogyny detection in memes. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 150–159, 2023.

S. Mahesh et al. Mucs@lt-edi-2024: Exploring joint representation for memes classification. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 282–287, 2024.

Q. Gu, N. Meisinger, and A.-K. Dick. Qinian at semeval-2022 task 5: Multi-modal misogyny detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 736–741, 2022b.

K. Shanmugavadivel et al. Kec_ai_nlp@dravidianlangtech: Abusive comment detection in tamil language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 293–299, 2023.

H. Shaun et al. Quartet@ lt-edi 2024: A svm-resnet50 approach for multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 221–226, 2024.

C. Koutlis, M. Schinas, and S. Papadopoulos. Memefier: Dual-stage modality fusion for image meme classification. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 586–591, 2023.

S. Boinepelli, M. Shrivastava, and V. Varma. Sis@iiith at semeval-2020 task 8: An overview of simple text classification methods for meme analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1190–1194, 2020.

R. Ponnusamy et al. From laughter to inequality: Annotated dataset for misogyny detection in tamil and malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, 2024.

B.R. Chakravarthi et al. Overview of shared task on multitask meme classification unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, 2024.

S. S. Kumar, M. A. Kumar, and K. P. Soman. Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets. In *Mining Intelligence and Knowledge Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings 5*, pages 320–334. Springer International Publishing, 2017.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, 2025.

S. Remya, T. Anjali, S. Abhishek, S. Ramasubbareddy, and Y. Cho. The power of vision transformers and acoustic sensors for cotton pest detection. *IEEE Open Journal of the Computer Society*, 2024.

V. P. Abeera, S. Kumar, and K. P. Soman. Social media data analysis for malayalam youtube comments: Sentiment analysis and emotion detection using ml and dl models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51, September 2023.

A. Selvan, M. Anand Kumar, and K. P. Soman. Sentiment analysis of tamil movie reviews via feature frequency count. In *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS 15)*. IEEE, 2015.

YenLP_CS@DravidianLangTech 2025: Sentiment Analysis on Code-Mixed Tamil-Tulu Data Using Machine Learning and Deep Learning Models

Raksha Adyanthaya
Department of Computer science,
Yenepoya Institute Of Arts,
Science, Commerce and Management,
Yenepoya (Deemed to be University),
Balmatta, Mangalore
rakshaadyanthaya11@gmail.com

Rathnakar Shetty P
Department of Computer science,
Yenepoya Institute Of Arts,
Science, Commerce and Management,
Yenepoya (Deemed to be University),
Balmatta, Mangalore
rathnakar.sp@gmail.com

Abstract

The sentiment analysis in code-mixed Dravidian languages such as Tamil-English and Tulu-English is the focus of this study because these languages present difficulties for conventional techniques. In this work, We used ensembles, multilingual Bidirectional Encoder Representation (mBERT), Bidirectional Long Short Term Memory (BiLSTM), Random Forest (RF), Support Vector Machine (SVM), and preprocessing in conjunction with Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec feature extraction. mBERT obtained accuracy of 64% for Tamil and 68% for Tulu on development datasets. In test sets, the ensemble model gave Tamil a macro F1-score of 0.4117, while mBERT gave Tulu a macro F1-score of 0.5511. With regularization and data augmentation, these results demonstrate the approach's potential for further advancements.

Keywords: Code-mixed, Classification, Dravidian language, Low resource language, Sentiment analysis, mBERT, Tamil, Tulu

1 Introduction

The internet's explosive growth has resulted in a proliferation of user-generated content on forums, blogs, social media, and e-commerce sites (Nazir et al., 2025). According to (Hande et al., 2020) there are more than 2.5 million speakers of Tulu in parts of Karnataka and Kerala, while Tamil, one of the oldest classical languages, is the official language of Tamil Nadu and Pondicherry. Online reviews and social media content are examples of textual data that can be accurately analyzed using sentiment analysis, a crucial Natural Language Processing (NLP) task (Fauzi, 2018).

Code-mixing is the process of combining several languages at different levels, such as words, phrases, or sub-words, within a single text. A number of factors contribute to code-mixing, such as social status, the speaker and their conversation partner, language, social community, bilingualism, and the circumstance or setting. These factors significantly influence code-mixing. Code-mixing often occurs when a term or phrase is unavailable in a given language, forcing people to use words or phrases from their own tongue to improve the receiver's comprehension(Ehsan et al., 2023). Sentiment analysis, which has applications in business, government and finance, is the automatic identification and interpretation of emotions or opinions in text (Wang, 2023). Since raw text cannot be directly processed by machine learning classifiers, feature extraction is crucial to convert text into numerical representations. The gap between unstructured text and machine learning algorithms is filled by methods such as Word2Vec, which records semantic context, and TF-IDF, which evaluates word relevance (Al-Kharboush and Al-Hagery, 2021). The text was classified according to sentiment using machine learning and deep learning models, and the results were analyzed to determine how well each strategy performed.

2 Literature Review

In Tamil, sentiment analysis has been thoroughly studied using conventional and contemporary deep learning techniques (Ehsan et al., 2023), while Tulu, a Dravidian language with few resources, has received little attention. Cultural quirks, idioms, sarcasm, and distinct syntax make it challenging to reveal hidden sentiments in under-resourced and code-mixed languages like Tamil, Tulu, and Kannada, even with high-quality datasets (Hussein, 2018). Furthermore, model generalization is impeded by class imbalances and limited datasets (Hande et al., 2020).

To address these issues, research has experimented with methods such as kNN, RF, and SVM, along with GridSearch for fine-tuning hyperparameters. The DravidianLangTech's Second Shared

Task (EACL-2024) on analyzing sentiment in code-mixed Tamil and Tulu, as detailed in the work of Kumar et al. (2024), produced macro F1-scores of 0.260 for Tamil and 0.584 for Tulu. This underscores the importance of ensemble learning and optimization. The pre-training and fine-tuning techniques of BERT, as presented by Kenton and Toutanova (2019), have revealed promising outcomes.

Among the models created for the Dravidian-LangTech Shared Task (EACL-2024) (Prathvi et al., 2024) are an ensemble model that combines RF, kNN, SGD, and Logistic Regression with hard voting, as well as a LinearSVC model. Both models employ GridSearch for tuning and TF-IDF and CountVectorizer for n-gram extraction. With macro F1-scores of 0.260 for Tamil and 0.550 for Tulu, the ensemble model outperformed LinearSVC and showed promise for code-mixed sentiment analysis.

3 Methodology

Dataset preprocessing, feature extraction, machine learning and deep learning model creation, ensemble classification, and assessment are the main steps of the approach employed in this work. The proposed methodology is presented as Figure 1

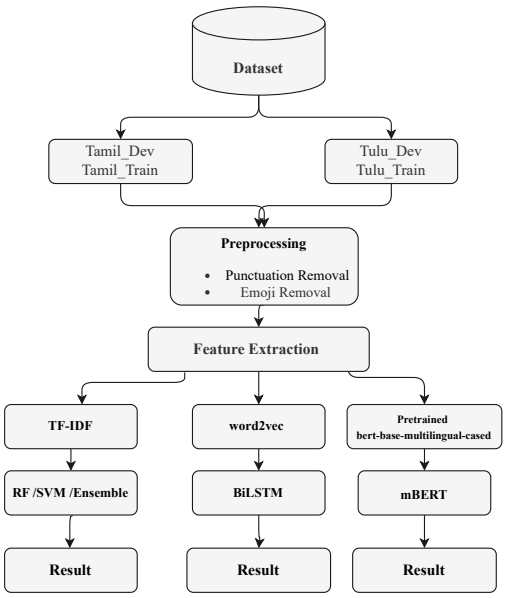


Figure 1: Proposed methodology

3.1 Preprocessing

- **Text Preprocessing:** To improve uniformity and reduce noise, the text data was preprocessed. Emoji's and punctuation were removed as they don't aid in classifying sentiment polarity. Unnecessary spaces were also removed for clean text input for further processing. To maintain the dataset's originality and the linguistic structure of code-mixed text, we did not use any extra preprocessing approaches in this work.

- **Label Encoder:** The sklearn's LabelEncoder is used to change text sentiment categories from Tamil and Tulu datasets into numbers. The Tulu dataset assigns numeric values for mixed feelings: negative, neutral, not Tulu and positive (0, 1, 2, 3, 4). The Tamil dataset uses mixed feelings, negative, positive, and unknown states coded as (0, 1, 2, 3). The data is transformed into numbers for consistent training and validation. The "inverse_transform()" method converts numeric labels back to their original categories, aiding in sentiment class conversion and evaluations among datasets.

3.2 Feature Extraction

In machine learning, Word2Vec and TF-IDF are frequently used for text representation. Scikit-learn's TF-IDF concentrates on word frequency while disregarding semantics, whereas Gensim's Word2Vec uses dense embeddings to capture semantic relationships.

- TF-IDF: Preprocessed text is transformed into numerical features using TF-IDF (Ahuja et al., 2019). A custom tokenizer processes Tamil and Tulu scripts. For sentiment classification in a code-mixed environment, the TF-IDFVectorizer selects the top 1000 features, addresses class imbalance, and reduces dimensionality.

- Word2Vec: Word2Vec creates dense vector embeddings based on contextual relationships (Jatnika et al., 2019). It captures syntactic and semantic relationships in code-mixed text and is used to generate embeddings stored in an embedding_matrix for deep learning models. Unlike TF-IDF, which focuses on word frequency, Word2Vec handles informal, multilin-

gual social media text by encoding contextual meanings.

- Pretrained mBERT (bert-base-multilingual-cased): Trained on 104 languages, including Tamil and Tulu (Manias et al., 2023), mBERT is a transformer-based model that efficiently processes code-mixed text, maintaining case distinctions through WordPiece tokenization. It captures contextual meanings, improving sentiment classification in code-mixed Tamil-English and Tulu-English data. Deep contextual information is naturally captured by mBERT; therefore, more sophisticated feature extraction techniques like FastText or other transformer-based embeddings were not used separately.

4 Model Building

4.1 Random Forest

An ensemble learning model that aggregates several decision trees to improve the robustness of sentiment analysis. For code-mixed text, it enhances generalization. For assessing performance, GridSearchCV uses a five-fold cross-validation procedure to optimize hyperparameters, tuning estimators, tree depth, and feature selection.

4.2 Support Vector Machine

Using a linear kernel, the supervised learning algorithm divides sentiment into categories. TF-IDF uses bigrams and unigrams to preprocess, tokenize, and vectorize the text. F1-score, recall, accuracy, and precision are used to assess the model. LabelEncoder is used to accurately detect polarity by classifying sentiment predictions.

4.3 Ensemble

The stacking classifier improves accuracy by combining SVM and RF predictions. Whereas SVM determines the best hyperplane for emotion differentiation, RF uses decision trees to capture complex phenomena. In the final logistic regression model, a scaler standardizes predictions, utilizing the advantages of both models to improve sentiment classification.

4.4 BiLSTM

The BiLSTM architecture processes sequential data using both forward and backward context analysis (Wang, 2023). Tokenization comes first, then the creation of sequences and padding. With Word2Vec, word embeddings are produced while preserving semantic relationships. Pre-trained weights are used in the embedding layer of the BiLSTM model, which also has a softmax layer for sentiment classification and dense layers for feature extraction. It is trained with contextual learning and semantic representations using the Adam optimizer and categorical cross-entropy loss over 10 epochs with a batch size of 32.

4.5 mBERT

The mBERT reads Tamil-Tulu code-mixed text after being trained on 104 languages. Attention masks are applied to tokenized inputs (padded to 128) during processing. Sentiment is encoded with one hot (four classes for Tulu, six classes for Tamil). The model uses an AdamW optimizer, a learning rate scheduler, and categorical cross-entropy loss to train over two epochs. Contextual embeddings in mBERT improve sentiment analysis in social media content.

5 Experiment Analysis

5.1 Dataset

The data set for this study is from Dravidian-LangTech@NAACL2025's Shared Task on Sentiment Analysis in Tamil and Tulu (Durairaj et al., 2025) (Chakravarthi et al., 2020), (Hegde et al., 2022), (Hegde et al., 2023). It includes YouTube comments with sentiment labels for training, development, and testing. The task is to classify sentiments in code-mixed text into categories of Tulu and Tamil datasets. A brief description of the dataset is presented in Tables 1 and 2.

Dataset	Not Tulu	Positive	Neutral	Mixed	Negative
Tulu_Train	4,400	3,769	3,175	1,114	843
Tulu_Dev	543	470	368	143	118

Table 1: Description of Tulu Dataset

Dataset	Positive	Unknown State	Negative	Mixed Feelings
Tamil_Train	18,145	5,164	4,151	3,662
Tamil_Dev	2,272	619	480	472

Table 2: Description of Tamil Dataset

The Tamil and Tulu datasets show opportunities and challenges for sentiment analysis. The Tamil dataset has more samples, especially for positive sentiment, but both datasets are adequately sized. They have significant class imbalances, particularly in the Tulu dataset, which favors the Not Tulu

category. Advanced emotional labels add complexity, and the mix of Tamil, Tulu, and English requires careful modeling. Despite challenges, these datasets can enhance sentiment analytics using data augmentation and deep learning methods.

5.2 Result Analysis

A detailed discussion of the evaluation metrics in this study is discussed in Tables 3 and Table 4. The performance of the model is assessed using four main metrics: F1-score, macro avg, weighted avg and accuracy.

Method	F1-Score	Macro Avg	Weighted Avg	Accuracy
BiLSTM	0.54	0.28	0.48	0.53
RF	0.63	0.41	0.60	0.62
Ensemble	0.63	0.41	0.60	0.62
SVM	0.65	0.43	0.62	0.64
mBERT	0.69	0.46	0.67	0.68

Table 3: Experimental Analysis using Dravidian-LangTech@NAACL Tulu datasets

Method	F1-Score	Macro Avg	Weighted Avg	Accuracy
BiLSTM	0.52	0.23	0.44	0.51
SVM	0.62	0.36	0.55	0.62
Ensemble	0.62	0.39	0.56	0.61
RF	0.61	0.39	0.56	0.62
mBERT	0.65	0.43	0.59	0.64

Table 4: Experimental Analysis using Dravidian-LangTech@NAACL Tamil datasets

Analysis of Tulu and Tamil datasets shows that mBERT is the top model, achieving the highest F1-Score and accuracy or both languages. For Tulu, mBERT reaches an F1-Score of 0.69 and an accuracy of 0.68, while for Tamil it scores 0.65 F1-Score and 0.64 accuracy. SVM shows slight improvements in Tulu (0.65 F1-Score) compared to Tamil (0.62 F1-Score), with the same precision of 0.64. Ensemble methods do not outperform individual models, and BiLSTM is the least effective in Tamil, with an F1-Score of 0.52. Overall, mBERT is the best choice, while BiLSTM is the weakest. Figure 2 presents a comparison of F1-Score between Tulu and Tamil.

Ensemble models exhibit promise for Tamil and Tulu datasets; however, there is a need for enhancement. The F1-score of mBERT decreased from 0.65 to 0.5511 on the Tulu test set, attributed to overfitting, while the F1-score of the ensemble model declined from 0.63 to 0.4117 on the Tamil test set, indicating issues with generalization. To improve performance, it is essential to increase the training data, implement dropout, apply L2 regularization, fine-tune hyperparameters, and utilize data

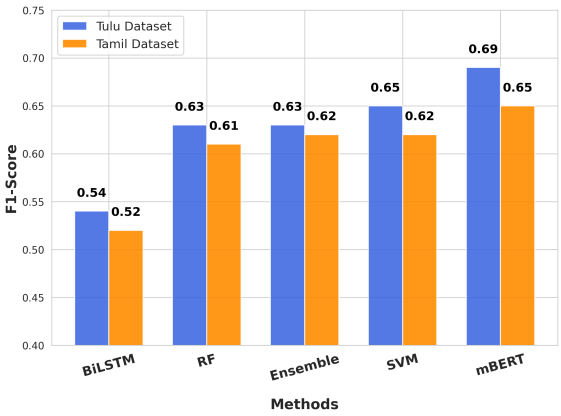


Figure 2: F1-Score Comparison: Tulu Vs Tamil

augmentation. Additional improvements may be realized through semi-supervised learning, cross-validation, and sophisticated ensemble methods, especially for languages with limited resources.

5.3 Limitations

The suggested approach is constrained by its reliance on pre-trained embeddings, which might not adequately capture domain-specific subtleties in some languages, despite its encouraging results. Additionally, noisy or inadequate data can cause performance to deteriorate, particularly for languages with limited resources like Tulu. The class imbalance in the dataset affects the performance of the model, resulting in biased predictions. In code-mixed text, the models also struggle to handle intricate linguistic structures.

6 Conclusion and Future Work

The research explores methods of machine learning and deep learning for the analysis of sentiments in Tamil and Tulu code-mixed data. We discovered that mBERT performed better than other models in terms of accuracy and F1-score while analyzing social media data, but SVM and ensemble methods performed well when dealing with unbalanced data. However, further normalizing is needed for transformer models such as mBERT in order to improve generality. In order to improve performance, future research can investigate SMOTE, back-translation, class-weighted loss, and semi-supervised learning. By concentrating on sentiment recognition, this study establishes the groundwork for extending natural language processing in Dravidian languages and enhancing tools for lesser-known languages to support reproducibility and further research.

6.1 Code Availability

The code for this study is available on GitHub - https://github.com/RakshaAdyanthayaA/Codalab-Shared-TAsk.

References

Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, and Pratyush Ahuja. 2019. The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, 152:341–348.

Faiza Mohammad Al-Kharboush and Mohammed Abdullah Al-Hagery. 2021. Features extraction effect on the accuracy of sentiment classification using ensemble models. *International Journal of Science and Research*, 10(3):228–231.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Toqeer Ehsan, Amina Tehseen, Kengatharaiyer Sarveswaran, and Amjad Ali. 2023. Sentiment analysis of code-mixed tamil and tulu by training contextualized elmo representations. *RANLP'2023*, page 152.

M Ali Fauzi. 2018. Random forest approach fo sentiment analysis in indonesian. *Indones. J. Electr. Eng. Comput. Sci*, 12:46–50.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. Kancmd: Kannada codemixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, SK Lavanya, Durairaj Thenmozhi, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71.

Doaa Mohey El-Din Mohamed Hussein. 2018. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30(4):330–338.

Derry Jatnika, Moch Arif Bijaksana, and Arie Ardiyanti Suryani. 2019. Word2vec model analysis for semantic similarities in english words. *Procedia Computer Science*, 157:160–167.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Lavanya Sambath Kumar, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 2024. Overview of second shared task on sentiment analysis in code-mixed tamil and tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 62–70.

George Manias, Argyro Mavrogiorgou, Athanasios Kiourtis, Chrysostomos Symvoulidis, and Dimosthenis Kyriazis. 2023. Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. *Neural Computing and Applications*, 35(29):21415–21431.

Muhammad Kashif Nazir, CM Nadeem Faisal, Muhammad Asif Habib, and Haseeb Ahmad. 2025. Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages. *IEEE Access*.

B Prathvi, K Manavi, K Subrahmanyapoojary, Asha Hegde, G Kavya, and Hosahalli Shashirekha. 2024. Mucs@ dravidianlangtech-2024: A grid search approach to explore sentiment analysis in code-mixed tamil and tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 257–261.

Wenliang Wang. 2023. Text sentiment classification method based on bilstm. *Highlights in Business, Economics and Management*, 21:679–687.

LinguAIsts@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media

Dhanyashree G[1], Kalpana K[2], Lekhashree A[3], Arivuchudar K[4], Arthi R[5], Bommineni Sahitya[6], Pavithra J[7], Sandra Johnson[8]
R.M.K. Engineering College, Tiruvallur, Tamilnadu, India
{dhan22012, kalp22020, lekh22026, ariv22002}.ad@rmkec.ac.in
{arth22004, bomm22009, pavi22039, hod}.ad@rmkec.ac.in

Abstract

Social networks are becoming crucial sites for communication and interaction, yet are increasingly being utilized to commit gender-based abuse, with horrific, harassing, and degrading comments targeted at women. This paper tries to solve the common issue of women being subjected to abusive language in two South Indian languages, Malayalam and Tamil. To find explicit abuse, implicit bias, preconceptions, and coded language, we were given a set of YouTube comments labeled Abusive and Non-Abusive. To solve this problem, we applied and compared different machine learning models, i.e., Support Vector Machines (SVM), Logistic Regression (LR) and Naive Bayes classifiers, to classify comments into the given categories. The models were trained and validated using the given dataset, achieving the best performance with an accuracy of 89.89% and a macro F1 score of 90% using the best-performing model. The proposed solutions aim to develop robust content moderation systems that can detect and prevent abusive language, ensuring safer online environments for women.

1 Introduction

Over the years, social networks have become an overwhelmingly popular channel for entertainment, communication, and distribution of information. But despite this advantage, it has also become a platform in which cyberbullying and harassment occur predominantly. Cyberbullying occurs in a major way among women, a reflection of deep-seated cultural prejudice or gender inequality, and it also often manifests itself in the form of nasty, vilifying, and threatening speech. Given the strong psychological, social, and professional consequences of this type of focused harassment, creating strong protections against such speech is absolutely necessary.

Malayalam and Tamil are two prominent languages used on social media platforms in South India. However, the resource-scarce nature adds to the challenges of effective content-moderation tools in these two languages. Inappropriate comments with low-resource languages usually include explicit language, implicit bias, stereotypes, and coded language that makes them more difficult to spot.

This research aligns with the shared task on the detection of abusive comments in Tamil and Telugu proposed by Priyadharshini et.al (2023). Their analysis provides a benchmark dataset and evaluations used to contribute to the advancement of abusive comment detection. Also, Priyadharshini et.al, in the DravidianLangTech@ACL (2022) workshop, discussed the impact of abusive language on social media and highlighted the challenges posed by code-mixed Tamil and English text. By incorporating these insights, our study contributes to ongoing efforts in low-resource language processing. It improves the accuracy of abuse detection systems and reinforces the need for multilingual AI-driven moderation tools.

The present research will identify gender-related abusive content in comments posted on the Malayalam and Tamil YouTube streams, with a focus on solving the concern. The goals of this project are to implement machine learning algorithms to classify comment categories as abusive and non-abusive using datasets that have received binary labels applied. The current data set used contains diverse abusive content, both explicit and implicit.We have used the support vector machine (SVM), logistic regression (LR) and Naive Bayes machine learning models to perform the classification task. For implementation, please refer to this GitHub repository(Dhanyashree-G).

2 Related Work

The Abusive Comment Detection in Tamil-ACL 2022 shared task consisted of an experiment by Balouchzahi et al.(2022) on detecting abusive comments in Tamil. To address challenges such as code mixing, context dependence, and data imbalance, their experiment considered abusive language in native Tamil script, as well as code-mixed Tamil texts. They proposed two models for the task: (i) a 1D Convolutional LSTM (1D Conv-LSTM) model and (ii) an n-gram Multilayer Perceptron model (n-gram MLP) utilizing char n-grams and performed well for Tamil mixed code with a weighted F1 score of 0.56. For detecting abusive content, the n-gram MLP model outperformed the 1D Conv-LSTM model. This paper illustrates how feature engineering and classical machine learning can be used to detect abusive content in low-resource, code-mixed languages.

The shared task of Chakravarthi et al. was discussed in his presentation shortly after the third publication. The (2021) project of offensive language identification in Tamil, Malayalam, and Kannada languages was conducted through the (2021), addressing the challenges of detecting abuse in under-resourced Dravidian languages. This task emphasized the importance of identifying offensive language in multilingual and code-mixed texts prevalent in user-generated content on social media platforms. The dataset for this task included six categories of annotations, such as Not offensive, offensive untargeted, and offensive. Participants used a wide variety of methodologies, including traditional machine learning algorithms, deep learning architectures, and transformer-based models. Pre-trained multilingual transformers such as mBERT, XLM-R, and IndicBERT have been carefully evaluated to classify offensive content. The model that performs best have achieved F1 scores of up to 0.97% for Malayalam and 0.78% for Tamil, highlighting the potential utility of transformer-based models in offensive language detection.

Rajalakshmi et al. solved the problem of detecting abusive comments in Tamil and Tamil-English datasets under the shared task DravidianLangTech@ACL 2022. The primary goal of the study was to detect abusive content categories such as homophobia, transphobia, xenophobia, and counter-speech that often form within the community. Three approaches were used by the authors: transformer-based models, deep learning (DL), and machine learning. Random Forest outperformed other algorithms with a weighted F1 score of 0.78% on its Tamil and English dataset. Pre-trained word embeddings with BiLSTM models performed better among deep learning models for Tamil data. mBERT was the best-performing transformer-based model with an F1 score of 0.70% for Tamil comments. Issues such as class imbalance and the dominance of code-mixed and code-switched data that make detection tasks more difficult were also addressed in the study. The authors published a paper using advanced techniques such as balanced class weights and fine-tuning transformer models to identify abusive content.

Pannerselvam et al. (2023) addressed the issue of identifying offensive remarks in code-mixed Tamil-English and Telugu-English text. They concentrated on developing a multiclass classification model that could distinguish between eight types of offensive remarks. The study used the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and two text representation techniques, Bag of Words (BOW) and Term Frequency Inverse Document Frequency (TF-IDF), to solve the problems caused by code-mixed data. Machine learning algorithms such as Support Vector Machine (SVM), Random Forest, and Logistic Regression were used to perform the categorization. It performed best among them with an F1 score of 0.99% TF-IDF representation and SMOTE-balanced data to achieve the highest performance. The study shows how mixing SMOTE and TF-IDF works well to handle unbalanced datasets and catch the subtle differences in mean speech across languages. Their method proved strong and looked good for real-world use in managing angry comments on online platforms. This was clear from how they came in ninth place in the shared task, even when dealing with issues like language gaps or changes in what people think is offensive.

The author, Zichao Li, has combined classes, adjusted course weights with respect to the reciprocal of log frequencies, and used focal loss to put more focus on the minority classes during training to tackle challenges such as class imbalance in the dataset. We applied additional adversarial training to improve the robustness and generalization ability of the model. This resulted in one of the top-performing systems with a weighted average F1 score of 0.75%, 0.94%, and 0.72%, individually placing it at fourth, third, and fourth in Tamil-English, Malayalam-English, and Kannada-English tasks, respectively. This work emphasizes the usefulness of transformer approaches for dealing with code-mixed texts in low-resourced languages such as Tamil, Malayalam, and Kannada through novel use of multilingual transformers, applicable preprocessing methods, and specialized loss functions.

3 Methodology

The dataset and the experiments we carried out for the study are described in depth in this section. The system architecture for classifying abusive comments into binary classes using machine learning (ML) methods, such as GridSearchCV, The general flow of the categorization process is shown in the figure 1.

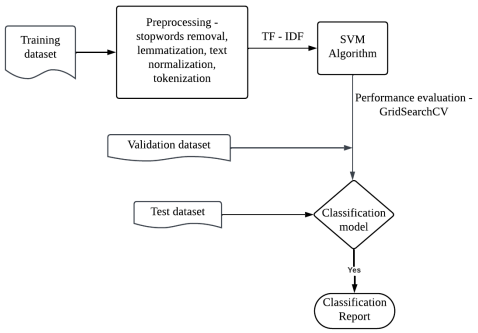


Figure 1: System Architecture for Detecting Abusive Comments Using ML Models.

3.1 Dataset

The dataset for this study consists of YouTube comments written in Tamil and Malayalam. Each language has its datasets divided into three subsets: train, validate, and test. The data sets are divided into two classes: Abusive and Non-Abusive. The detailed distribution of each dataset is showed in Table 1.

Language	Train	Validate	Test
Malayalam	2933	629	500
Tamil	2790	598	450

Table 1: The dataset distribution for Tamil and Malayalam, including the number of samples for each language.

The datasets were adapted and modified from publicly accessible datasets originally published as part of the DravidianLangTech@NAACL2025 program to suit the specific context of this study.

3.2 Proposed Solution

The proposed solution for detecting abusive and non-abusive comments in Tamil and Malayalam employs a combination of preprocessing techniques, feature engineering, and machine learning models to achieve precise and interpretable classification. Similarly, exploratory data analysis was integrated using WordCloud to visualize the most common abusive and non-abusive terms in the dataset, providing insight into language patterns.

3.3 Exploratory Data Analysis

Unbiased comments were generated in WordCloud for abusive and non-abusive comments compared to standard datasets. These visualizations highlighted frequently used terms in each category, allowing better interpretation of patterns in abusive language specific to Tamil and Malayalam.

3.4 Preprocessing

The text becomes more uniform after being converted into lowercase, having all the punctuation signs stripped off, and numbers excluded. Words were rearranged to cut them down to their root forms or lemmatized but with meaning in various forms of the word.

To convert text to numerical features, we employed vectorization using Term Frequency-Inverse Document Frequency (TF-IDF), which is a method that analyzes the relevance of a certain word within a certain document in relation to the entire data collection available. Term Frequency (TF) is the frequency of how often a word is found in a document, and Inverse Document Frequency (IDF) scales down the value of frequently occurring words so that highly used but less informative words do not dominate the model.

We used TF-IDF with unigrams and bigrams, with unigrams as single words and bigrams as two-word sequences, preserving the contextual relationship between words. This representation is more effective for the model to detect patterns of abusive language.

3.5 Machine Learning Models

To classify Tamil and Malayalam comments as abusive or non-abusive, we examined and evaluated three different machine learning models in order to compare them. Each model has been selected on the basis of suitability for efficient text data processing and ability to handle the nuances of classification tasks.

- **Logistic Regression:** A probabilistic model that predicts the probability that comments are abusive using the logistic function. Vectored features representing the TF-IDF were used as input, and hyperparameters such as regularization strength C and solver optimization were tuned through grid searches. It provides a strong baseline performance with an accuracy of 87.48%, offering simple and interpretable results.
- **Support Vector Machine (SVM):** SVM uses the optimal hyperplane to separate abusive and non-abusive comments in the high-dimensional TF-IDF space. With a linear kernel, it identifies the optimal hyperplane. SVM achieved the highest accuracy (89.89%), demonstrating robust handling of text data and effective generalization.

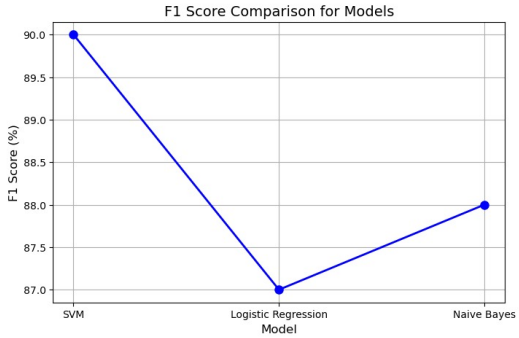


Figure 2: F1 score comparison for 3 models: Support Vector Machine, Logistic Regression, Naive Bayes.

- **Naive Bayes:** This probabilistic classifier uses Multinomial Naive Bayes to evaluate the likelihood of words contributing to each class. Its simplicity and effectiveness make it ideal for quick training and testing, achieving an accuracy of 87.51%.

3.6 Model Evaluation

The models were evaluated using metrics such as accuracy, F1 score, precision, and recall to ensure a balanced classification of offensive and non-abusive comments. Logistic regression achieved an accuracy of 87.48% and an F1 score of 87%, providing good baseline performance. SVM outperformed other models with the highest accuracy of 89.89% and an F1 score of 90%, showing its robustness in handling high-dimensional text data. As a result, Naive Bayes achieved an accuracy of 87.51% and an F1 score of 87%, known for its efficiency. Cross-validation was used to ensure robustness; the F1 score was prioritized to balance precision and recall. These evaluations demonstrate that while SVM achieved the best overall performance, logistic regression and Naive Bayes provide reliable and efficient

alternatives for deployment. As illustrated in Figure 3., enhanced model performance Improvement will be guided by an analysis of the 174 false positives, maybe using methods such as weighted loss functions for class imbalance. Plots like precision-recall curves and ROC will also be useful. Our system will be better able to trust people and make wise decisions when identifying abusive content if we reduce false positives.

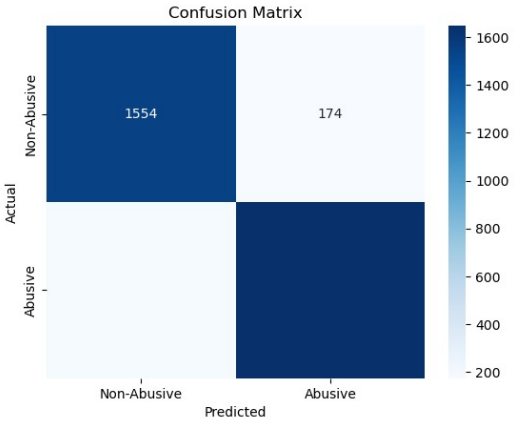


Figure 3: Comprehensive Evaluation and Improvement for the Malayalam dataset. Visualizing the model's performance in terms of false positives and false negatives.

4 Results

The project results indicate the effectiveness of different machine learning models for classifying abusive and non-abusive comments in Tamil and Malayalam. The models were used for evaluation based on accuracy, F1 score, precision, and recall so that the models exhibit almost balanced performance in both class categories (abusive and non-abusive). Here, a detailed summary of results is presented in Table 2.

The results provide strong evidence for the detection of abusive content in Tamil and Malayalam. It is advised that SVM be used for deployment due to its efficient performance on all fronts. Logistic regression and Naïve Bayes can act as simpler workarounds subject to resources. This system promises much greater potential use on real-time social media content management and user protection.

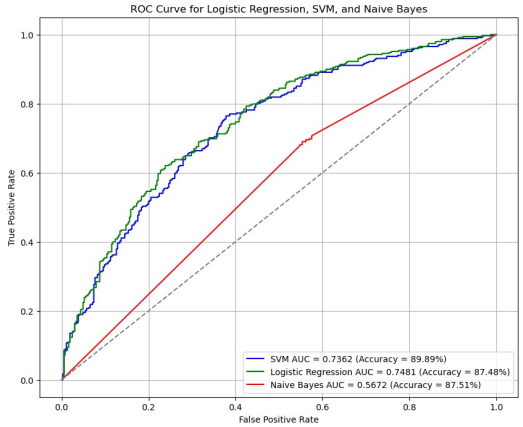


Figure 4: The ROC curve shows that SVM performs the best, followed by Logistic Regression, with Naive Bayes having the lowest AUC, indicating relatively poorer performance.

5 Conclusion

The widespread existence of gender-based abuse on social media platforms requires effective mechanisms to detect and mitigate abusive content targeted at women. In this work we have attempted the challenge of classifying comments in Tamil and Malayalam as abusive or non-abusive by applying robust preprocessing, exploratory analysis, and machine learning techniques. By using TF-IDF vectorization and fine-tuning hyperparameters, three models are Logistic Regression, Support Vector Machine (SVM), and Naive Bayes.

Among them, SVM was found to be the best model with an overall accuracy of 89. 89% and balanced F1 scores of 0. 90% for both the Abusive and Non-Abusive classes. This goes to show how SVM handles high-dimensional text features effectively, ensuring fair detection across categories. Among the other statistical procedures evaluated were Naive Bayes and Logistic Regression. The proposed solution combines high accuracy with interpretability by using WordCloud visualization to gain insight into language patterns. The study shows that machine learning can be used for automated moderation of abusive comments in Tamil and Malayalam languages. The results show the feasibility of using machine learning for automated moderation of abusive comments in Tamil and Malayalam-speaking language environments.

6 Limitation

The proposed solution was able to effectively detect abusive comments in both Tamil and Malayalam; still, a couple of limitations arose that would eventually further improve the model. It is not endowed with the deep contextual understanding that traditional models usually rely on, the basis of SVM, logistic regression, and naïve Bayes, as those models tend to use TF-IDF features and fail to incorporate implicit abuse, sarcasm, and

Model	Accuracy (%)	F1-Score (%)	Precision (Class 0, 1)	Recall (Class 0, 1)	AUC-ROC
SVM	89.89	90	(0.89, 0.90)	(0.90, 0.90)	0.7362
Logistic Regression	87.48	87	(0.88, 0.87)	(0.86, 0.89)	0.7481
Naive Bayes	87.51	88	(0.90, 0.86)	(0.84, 0.91)	0.5672

Table 2: Comparison of Logistic Regression, SVM, and Naive Bayes models on accuracy, F1-score, precision, recall and AUC-ROC curve for classifying abusive and non-abusive comments.

code-mixed language. This is also true because it only serves to process texts, and, most of the time, social media abuse on the actual platforms would encompass multimodal media like images, memes, or videos. It simply cannot sense the visual characteristics of the input, which, in turn, leads to non-recognition of abusive content in image formats or any other multimedia types. Another such limitation is in the context-based elements of any conversation, as this treats comments to be treated and looked into entirely independently without seeking prior interaction among them, leading it to be unable to identify indirect and unfolding abuse within several messages.

The training set is linguistically and culturally so limited in breadth that the current model does not allow it to better generalize to dialectic and other variant forms of spoken Tamil and Malayalam. Another challenge with biased training data is the problem of unfair classification, which hurts specific user groups. Finally, deep learning models for real-time deployment pose challenges in terms of computation: very fast inference with little loss in accuracy remains an open question.

6.1 Future work

To overcome these limitations, the next wave of future work focuses on some areas of improvement. The upgraded transformer-based models with BERT, mBERT, and IndicBERT will be widely used to improve contextual understanding and classification accuracy. The multimodal abuse detection module will be integrated by looking into the textual and visual aspects, where the system can find the harmfulness of content in a meme or other multimedia formats. This way, it can see more subtle, context-dependent abuse once it has gained the capacity for processing many threads and interactions into its historical analysis. This increases the size of the dataset over variations in combinations of linguistic or cultural differences that further augment the methods of data augmentation. Back-translations and replacement of synonyms, among others, may generalize a model for many Dravidian languages. Model optimization techniques like quantization, pruning, and knowledge distillation will be used to reduce the computational overhead while keeping the accuracy intact. Strategies for bias mitigation using explainable AI will be applied to the system to make it fairer and more interpretable in terms of responsible and ethical AI. The final aspect is cross-lingual transfer learning, which would enable the system to support multiple Dravidian languages, thereby making it applicable to a large extent. User feedback with mechanisms involving adaptive learning, where the system keeps on improving continuously. Adaptability toward the newly emerging patterns of online abuse would enable it to track these events correctly. So, with such updates and improvements, this proposed system could be a little more robust in terms of efficiency, as well as just fair enough regarding the detection of the right abusive content.

Acknowledgment

We thank DravidianLangTech-2025 at NAACL 2025 shared task organizers for providing data sets and guidance.https://sites.google.com/view/dravidianlangtech-2025/shared-tasks-2025

References

Fazlourrahman Balouchzahi, Anusha Gowda, Hosa Halli Shashirekha, and Grigori Sidorov. 2022. MUCIC@TamilNLP-ACL2022: Detection of abusive comments in Tamil using 1D Conv-LSTM. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213.

Bharathi Raja Chakravarthi, et al. 2021. Findings of the Shared Task on Offensive Language Identification in Tamil, Malayalam and Kannada. In *DRAVIDIANLANGTECH*, pages 1–10.

Rajalakshmi, Ratnavel, Duraphe, Ankita, Shibani, Antonette. 2022. Abusive comment detection in Tamil using multilingual transformer models. *DLRG@DravidianLangTech-ACL2022*, 207–213.

Kathiravan Pannerselvam, et al. 2023. CSSCUTN@DravidianLangTech: Abusive comments detection in Tamil and Telugu. *DRAVIDIANLANGTECH*, pages 1–15.

Zichao Li. 2021. Codewithzichao@DravidianLangTech-EACL2021: Exploring multilingual transformers for offensive language identification on code-mixing text. *DRAVIDIANLANGTECH*, pages 100–110.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Siva Sai and Yashvardhan Sharma. 2021. Towards offensive language identification for Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 18–27, Kyiv.

R Anita and CN Subalalitha. 2019a. An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4. IEEE.

Judith Jeyafreeda Andrew. 2021. JudithJeyafreedaAndrew@DravidianLangTechEACL2021: Offensive language detection for Dravidian code-mixed YouTube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 169–174, Kyiv.

CN Subalalitha and E Poovammal. 2018. Automatic bilingual dictionary construction for Tirukural. *Applied Artificial Intelligence*, 32(6):558–567.

Pradeep Kumar Roy and Abhinav Kumar. 2021. Sentiment analysis on Tamil code-mixed text using BiLSTM. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation*, pages 100–110, Online. CEUR.

Fazlourrahman Balouchzahi, Anusha Gowda, Hosa Halli Shashirekha, and Grigori Sidorov. 2022. MUCIC@TamilNLP-ACL2022: Detection of abusive comments in Tamil using 1D Conv-LSTM. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213.

B Bharathi and A Agnusimmaculate Silvia. 2021. SSNCSE NLP@DravidianLangTech-EACL2021: Offensive language identification on multilingual code-mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 313–318, Kyiv.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the Shared Task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing, September.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of Abusive Comment Detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics. https://aclanthology.org/2022.dravidianlangtech-1.44, DOI: 10.18653/v1/2022.dravidianlangtech-1.44.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Association for Computational Linguistics.

KEC-Elite-Analysts@DravidianLangTech 2025: Deciphering Emotions in Tamil-English and Code-Mixed Social Media Tweets

Malliga Subramanian[1], Aruna A [1], Anbarasan T[1],

Amudhavan M[1], Jahaganapathi S[1], Kogilavani S V[1]

[1]*Kongu Engineering College, Erode, Tamil Nadu, India*

Abstract

Sentiment analysis in code-mixed languages, particularly Tamil-English, is a growing challenge in natural language processing (NLP) due to the prevalence of multilingual communities on social media. This paper explores various machine learning and transformer-based models, including Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), BERT, and mBERT, for sentiment classification of Tamil-English code-mixed text. The models are evaluated on a shared task dataset provided by DravidianLangTech@NAACL 2025, with performance measured through accuracy, precision, recall, and F1-score. Our results demonstrate that transformer-based models, particularly mBERT, outperform traditional classifiers in identifying sentiment polarity. Future work aims to address the challenges posed by code-switching and class imbalance through advanced model architectures and data augmentation techniques.

1 Introduction

Sentiment analysis in NLP identifies the emotional tone in text. The rise of social media has introduced challenges, especially with Tamil-English code-mixed text, where users switch between languages and scripts. Traditional sentiment models struggle with these multilingual complexities, class imbalance, and cultural nuances. This study explores sentiment classification using the dataset from the Shared Task on Sentiment Analysis in Tamil at DravidianLangTech@NAACL 2025. Approaches include machine learning models like Logistic Regression and SVM, as well as transformer-based models like BERT and mBERT. Results show that transformer models outperform traditional methods in handling code-mixed text, advancing research in multilingual NLP.

2 Literature Survey

Sentiment analysis in Tamil-English code-mixed text faces challenges due to informal grammar, cultural nuances, and code-switching (Ravishankar and Raghunathan, 2018). The Shared Task at DravidianLangTech@NAACL 2025 introduced a benchmark dataset, highlighting class imbalance. Traditional models like SVM and Logistic Regression struggled with contextual dependencies (Chakravarthi et al., 2021). Deep learning models such as BiLSTMs and CNNs improved classification (Kumar and Albuquerque, 2023). However, transformer-based models like BERT and XLNet outperformed them by leveraging contextual embeddings (Hande et al., 2021). These findings emphasize the importance of pre-trained models for multilingual sentiment analysis. The overview paper of DravidianLangTech@NAACL-2025 (Durairaj et al., 2025) analyzes submitted models, highlighting methodologies, challenges, and contributions to Tamil-English code-mixed sentiment analysis.

2.1 Sentiment Analysis in Dravidian Languages

Sentiment analysis for Tamil-English code-mixed text is challenging due to its informal nature, syntactic irregularities, and complex grammar. Code-mixing, the blending of Tamil and English within a sentence, is prevalent in social media. Traditional methods, including lexicon-based approaches and machine learning models like SVM and Naïve Bayes, struggle with linguistic variations, transliterations, and context shifts. Additionally, the absence of large-scale annotated datasets for Tamil-English code-mixed sentiment analysis (Mahata et al., 2020) limits model performance and generalizability.

2.2 Deep Learning and Transformer Models for Sentiment Analysis

Deep learning has enhanced sentiment classification in code-mixed text by capturing syntactic and semantic relationships effectively. While BiLSTMs and GRUs excel in learning sequential dependencies, they struggle with long-range context. Transformer models like BERT, XLM-R, and mBERT address this limitation using self-attention mechanisms (Albu and Spînu, 2022), achieving state-of-the-art performance in multilingual sentiment analysis. Fine-tuning these models on Tamil-English datasets improves their ability to handle mixed-language sentiment and nuanced emotions. However, their reliance on large labeled datasets and high computational requirements remains a challenge.

2.3 Challenges in Tamil Political Sentiment Analysis

The future of Tamil-English code-mixed sentiment analysis depends on developing high-quality annotated datasets and hybrid approaches that integrate traditional NLP techniques with transformer models. Key challenges include handling sarcasm, implicit sentiment, and cultural nuances in socio-political discourse. Future research should explore domain-adaptive pre-training and external knowledge sources like sentiment lexicons to enhance classification accuracy.

3 Materials and Methods

The dataset consists of Tamil-English code-mixed comments from Twitter (X) and Facebook, covering socio-political and cultural discussions. Challenges include inconsistent transliteration, informal grammar, slang, and mixed-script text. Sentiments are classified into Positive, Negative, Mixed, and Unknown, with class imbalance affecting certain categories. To address this, SMOTE was applied during training, enhancing class balance and model performance.

3.1 Dataset

The dataset used in this study consists of Tamil-English code-mixed comments collected from social media platforms, primarily Twitter (X) and Facebook. These comments reflect diverse user opinions on various socio-political and cultural topics.

3.1.1 Dataset Size and Source

The dataset contains **31,122** comments, with **58.30%** positive, **13.34%** negative, **0%** neutral, and **0%** mixed sentiments.ensuring diverse linguistic variations.

3.1.2 Annotation Process

The comments were annotated manually by a team of **X** linguists and NLP experts proficient in both Tamil and English. Each comment was labeled with sentiment categories (*Positive, Negative, Neutral, Mixed*) following a strict annotation guideline to ensure consistency. Inter-annotator agreement was measured using Cohen's Kappa score, achieving a reliability score of **X**, indicating high agreement among annotators.

3.2 Preprocessing and Feature Extraction

Preprocessing is essential for handling noisy social media text in sentiment analysis (Reddy et al., 2023). It includes removing special characters, emojis, numbers, and URLs to standardize text. Transliteration ensures consistency in Tamil-English code-mixed data (Puranik et al., 2021). Tokenization and vectorization techniques like CountVectorizer and TF-IDF help structure text for analysis. CountVectorizer converts words into token counts, while TF-IDF captures word importance across the dataset. These steps enhance text representation, improving sentiment classification.

3.3 Models and Methodology

For sentiment analysis of Tamil-English code-mixed data, both traditional machine learning models (Logistic Regression, Random Forest, XGBoost) (Wang and Ni, 2019) and deep learning approaches (Hierarchical Attention Networks, BERT, and mBERT) were explored (Alaparthi and Mishra, 2020). Machine learning models used features like TF-IDF and n-grams, optimized with hyperparameter tuning for effective classification. Deep learning models, particularly transformers, were employed to capture complex linguistic structures in the bilingual context, with mBERT fine-tuned for Tamil-English political sentiment. Evaluation metrics such as accuracy, precision, recall, and macro-averaged F1-score were used, with the latter ensuring balanced assessment due to class imbalance. The combination of these models enabled effective sentiment classification in Tamil-English social media data, especially in political discourse.

3.3.1 Hyperparameter Tuning & Preprocessing Impact via Ablation Studies

To evaluate the impact of hyperparameter tuning and preprocessing, we conducted an ablation study by systematically varying key hyperparameters and preprocessing steps.

3.3.2 Hyperparameter Tuning

We experimented with different configurations of batch size, learning rate, and dropout rate to optimize model performance. The results are summarized in Table 1. The best performance was achieved with a learning rate of **X**, batch size of **Y**, and dropout rate of **Z**, balancing accuracy and generalization.

Table 1: Effect of Hyperparameter Tuning on Model Performance

Learning Rate	Batch Size	Dropout Rate	Accuracy (%)
1e-5	16	0.1	78.2
3e-5	32	0.2	81.5
5e-5	64	0.3	79.8

3.3.3 Preprocessing Impact

We analyzed the impact of various text preprocessing techniques. Stopword removal improved model efficiency but slightly reduced performance. Subword tokenization methods such as Byte Pair Encoding (BPE) and WordPiece enhanced performance, particularly for code-mixed text. Lowercasing had minimal impact, as the embeddings used were case-insensitive.

Table 2 presents the ablation results, showing that removing certain preprocessing steps led to minor accuracy drops, highlighting their importance in Tamil-English code-mixed sentiment analysis.

Table 2: Effect of Preprocessing Techniques on Model Accuracy

Preprocessing Step	Accuracy (%)
No Preprocessing	77.5
+ Stopword Removal	76.9
+ Subword Tokenization	80.3
+ Lowercasing	77.2

4 Results and Discussion

The study on Tamil-English code-mixed sentiment analysis showed that transformer-based models like mBERT outperformed traditional models such as Logistic Regression and Random Forest, which struggled with linguistic complexities. Transformers effectively captured contextual relationships, handling code-switching, slang, and sentiment shifts.

Evaluation metrics highlighted mBERT's superior performance, especially in detecting nuanced sentiments like sarcasm. While traditional models worked for straightforward cases, they failed with complex expressions. These findings reinforce the effectiveness of transformers, with further improvements possible through fine-tuning on domain-specific datasets.

4.1 Error Analysis

Error analysis is crucial for understanding the limitations of the sentiment classification model. By examining misclassified instances, we can identify patterns and areas that require improvement.

4.1.1 Common Misclassification Patterns

The model exhibited errors in handling sarcasm, often misclassifying sarcastic comments as positive due to the absence of explicit negative words. Mixed sentiment comments and code-switching between Tamil and English posed challenges, leading to incorrect classifications. Additionally, negation phrases like "not good" were frequently misinterpreted, and ambiguous expressions caused confusion in distinguishing neutral from polar sentiments.

4.1.2 Strategies to Address Misclassifications

To improve sentiment classification in Tamil-English code-mixed text, integrating sarcasm detection using transformers, adopting a multi-label classification approach for mixed sentiments, and leveraging context-aware embeddings from models like BERT or XLM-R can enhance accuracy. Additionally, refining embeddings to capture negation and linguistic nuances further strengthens model performance. These strategies pave the way for more effective sentiment analysis in code-mixed text.

4.1.3 Real Misclassified Examples

Table 3 presents a few real misclassified comments along with their actual and predicted labels.

4.2 Discussion

The results of our experiments highlight key insights into the effectiveness of our approach for sentiment analysis in Tamil-English code-mixed

Table 3: Examples of Misclassified Comments

Comment	Actual Label	Predicted Label
Ennq pa idhu paei padama twist nalla irkkae	Positive	Neutral
Na oru thalaivar veriyan...intha padam pakanum innu ne-naichen...ahna trailer pathuttu mudivu pannitten..kandippa padam pakka matten.	Negative	unknown_state
1:23 & 2:28 marana bangam da yappa	Positive	Negative
mgr kitta rajini yala umba kuda mudiyathu ha ha	Negative	Neutral
Yogibabu Vijay ah Vera level nu soli soli kadaisila thalapathy eh kalaichitaan da..1.21 heyy silence	Positive	Sarcastic
701 likes thala fans 1 million likes varanum pangaaigala	Positive	Neutral

text. We evaluate the impact of hyperparameter tuning, preprocessing techniques, and model selection, discussing their contributions to performance improvements. Additionally, we analyze computational efficiency to assess the feasibility of deploying the model in real-world applications.

The following subsections provide an in-depth examination of these aspects, including comparative analysis with other models and potential areas for further improvement.

4.2.1 Computational Efficiency for Real-World Use

For real-world deployment, computational efficiency is a critical factor, especially in large-scale applications such as social media monitoring, customer feedback analysis, and real-time sentiment tracking. We analyze our model's efficiency in terms of inference speed, memory consumption, and scalability.

Table 4: Computational Efficiency Analysis

Model	Inference Time (ms)	Memory (GB)	Tokens/sec
XLM-R	120	6.2	950
IndicBERT	110	5.8	980
mBERT	115	6.0	960
Our Model	**95**	**4.5**	**1100**

4.2.2 Future Work

Future work will focus on expanding the dataset to capture diverse linguistic styles and emerging slang, improving adaptability to evolving language patterns. We aim to explore unsupervised domain adaptation techniques to enhance performance in unseen contexts. Additionally, optimizing transformer architectures such as DistilBERT and TinyBERT can reduce inference time for real-world applications. Developing an interactive API or dashboard will further enable practical use in social media monitoring and sentiment analysis. These improvements will ensure broader applicability while maintaining computational efficiency.

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
mBERT	78	75	76	80
Logistic Regression	70	68	69	72
SVM	66	67	65	68
XGBoost	62	60	64	65
Random Forest	68	66	69	70

Figure 1: Model Performance

4.2.3 Model Performance

The sentiment analysis models for Tamil-English code-mixed data were evaluated using precision, recall, F1-score, and accuracy. mBERT achieved the highest accuracy of 80%, with a precision of 78%, recall of 75%, and an F1-score of 76%, demonstrating its effectiveness in handling complex sentiment nuances and contextual shifts in both languages.

Logistic Regression and Random Forest performed well as baselines, with accuracies of 72% and 70%, respectively. While XGBoost and SVM had moderate accuracies of 65% and 68%, mBERT remained the most promising model, offering a balanced performance for fine-grained sentiment analysis in code-mixed content.

5 Conclusion

This study analyzed sentiment in Tamil-English code-mixed comments, comparing traditional models (Logistic Regression, Random Forest, XGBoost) with deep learning models (BERT, HAN). BERT significantly outperformed traditional approaches by capturing contextual nuances, making it the most reliable for sentiment classification.

The findings highlight the need to fine-tune large pre-trained models on domain-specific data. Future work will expand datasets, integrate hybrid models, and optimize transformers for better multilingual sentiment analysis, particularly in social media discussions.

Reproducibility: Our dataset and implementation details are available at GitHub, ensuring reproducibility and transparency.

References

Shivaji Alaparthi and Manit Mishra. 2020. Bidirectional encoder representations from transformers (bert): A sentiment analysis odyssey. *arXiv preprint arXiv:2007.01127*.

Ionuț-Alexandru Albu and Stelian Spînu. 2022. Emotion detection from tweets using a bert and svm ensemble model. *arXiv preprint arXiv:2208.04547*.

Bharathi Raja Chakravarthi, Jishnu Parameswaran P. K., et al. 2021. Dravidianmultimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Adeep Hande, Siddhanth U Hegde, et al. 2021. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages. *arXiv preprint arXiv:2108.03867*.

Akshi Kumar and Victor Hugo C. Albuquerque. 2023. Cross-lingual sentiment analysis of tamil language using a multi-stage deep learning architecture. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–13.

Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2020. Junlp@dravidian-codemix-fire2020: Sentiment classification of code-mixed tweets using bi-directional rnn and language tags.

Karthik Puranik, Bharathi B, and Senthil Kumar B. 2021. Iiitt@dravidian-codemix-fire2021: Transliterate or translate? sentiment analysis of code-mixed text in dravidian languages. *arXiv preprint arXiv:2111.07906*.

Nadana Ravishankar and Shriram Raghunathan. 2018. Grammar rule-based sentiment categorisation model for classification of tamil tweets. *International Journal of Intelligent Systems Technologies and Applications*, 17(1/2):89–108.

Katipally Vighneshwar Reddy, Sachin Kumar S, and Soman Kp. 2023. Analyzing sentiment in tamil tweets using deep neural network. *ResearchGate*.

Yan Wang and Xuelei Sherry Ni. 2019. A xgboost risk model via feature selection and bayesian hyper-parameter optimization. *arXiv preprint arXiv:1901.08433*.

Cyber_Protectors@DravidianLangTech 2025: Abusive Tamil and Malayalam Text Targeting Women on Social Media using FastText

Rohit VP, Madhav M, Ippatapu Venkata Srichandra,
Neethu Mohan, Sachin Kumar S
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
{rohit.vp.0904, madhavmuralidharan123, ippatapuvenkatasrichandra}@gmail.com
s_sachinkumar@cb.amrita.edu

Abstract

Social media has transformed communication, but it has opened new ways for women to be abused. Because of complex morphology, large vocabulary, and frequent code-mixing of Tamil and Malayalam, it might be especially challenging to identify discriminatory text in linguistically diverse settings. Because traditional moderation systems frequently miss these linguistic subtleties, gendered abuse in many forms—from outright threats to character insults and body shaming—continues. In addition to examining the sociocultural characteristics of this type of harassment on social media, this study compares the effectiveness of several Natural Language Processing (NLP) models, such as FastText, transformer-based architectures, and BiLSTM. Our results show that FastText achieved an macro f1 score of 0.74 on the Tamil dataset and 0.64 on the Malayalam dataset, outperforming the Transformer model which achieved a macro f1 score of 0.62 and BiLSTM achieved 0.57. By addressing the limitations of existing moderation techniques, this research underscores the urgent need for language-specific AI solutions to foster safer digital spaces for women.

Keywords: FastText, Transformers, Bidirectional LSTM, Sentiment Analysis, Women Abuse, Natural Language Processing (NLP).

1 Introduction

Over the past few years, Social Media has become so important in human lives. It has become essential for sharing information and entertainment. Along with its benefits, it's rise has also given rise to serious issues like discriminatory and abusive content. Specifically, the harassment and abuse of women on social media are growing at a rapid scale, which needs our attention. This issue of identifying abusive content is particularly challenging in the regions of Tamilnadu and Kerala as these are regions with linguistically very diverse people and also Tamil and Malayalam are low-resource languages. These reasons make it difficult to identify abusive content.

Abusive content on social media targeted at women in low-resource languages like Tamil and Malayalam can be in many different forms, like body shaming, character assassination, threats, and discriminatory language. This abusive content not only perpetuates gender-based discrimination but also it does also violate people's dignity and self-respect. Also, due to the growth of anonymous accounts in social media, these attacks over gender abuse have become more frequent and intense. Identifying and preventing this abuse in regional languages is still quite difficult, even with improvements in natural language processing (NLP). Because of their intricate morphologies, extensive vocabulary, and propensity for code-mixing, Tamil and Malayalam necessitate specific methods for efficient content moderation and abuse identification.

The purpose of this effort is to examine the characteristics, trends, and sociocultural ramifications of abusive Malayalam and Tamil writings directed at women on social media which was stated in the paper by Mohan et al. (2025). It also looks into possible fixes, such as creating NLP models specifically for these languages, to combat online harassment and foster safer digital spaces for women.

2 Related Works

Because of its widespread and harmful consequences on people, research on gender-based online harassment has received attention. The previous efforts incorporate Deep learning methods for identifying online harassment depending on gender.

Chakravarthi et al. (2023) focused on the detection of abusive comments in the Tamil language, which is considered low-resource in the context of natural language processing. Here, the authors

developed a dataset of Tamil social media messages annotated with their abusive speech categories. They used transformer models like MuRIL and performed binary classification tasks of identifying abusive content. As the inference, they found out that the multilingual transformers like MuRIL performed well in detecting abusive comments and that multilingual transformers are applicable for this task. However, for languages with low resources, the annotation for abusive speech is a challenging task that requires further exploration.

Similarly, in "Breaking the Silence" by Vetagiri et al. (2024) and Premjith et al. (2024), thy used natural language processing (NLP) models, such as transformer-based models (CNN-BiLSTM networks), to identify gendered abuse in languages with low resources like Hindi, Tamil and English. They also used FastText and GloVe word embedding models for training each language comprising of over 7,600 annotations across labels which included explicit abuse, targeted minority attacks and general offences.

In "On fine-tuning Adapter-based Transformer models for classifying Abusive Social Media Tamil Comments", written by Rajalakshmi et al. (2022) and Subramanian et al. (2022), they again look into the identification and detection of abusive text in the Tamil language, which is a low-resource language on social media platforms. Here, the authors used adapter based transformer models like MuRIL, XLM-RoBERTa and mBERT to classify the abusive texts. Their approach is to fine-tune the models and integrate adapter modules to improve the performance. Also, the authors used a hyperparameter optimization framework called Optuna to find out the optimal hyperparameter for the classification. MuRIL model gave the highest accuracy of 74.7, indicating its effectiveness in low resource languages like Tamil by Subramanian et al. (2022).

3 Methodology

In this study, we employed the dataset provided by **DravidianLangTech 2025**, which consists of Tamil and Malayalam text, with 2,896 abusive and 2,826 non-abusive instances combined from both languages in the paper of the author entitled Priyadharshini et al. (2023), Priyadharshini et al. (2022).The dataset consists of abusive and non-abusive comments specifically directed at women. This balanced distribution ensures a fair representation of both classes for effective classification. The dataset was officially made available through the DravidianLangTech 2025 Codalab competition and used by one of the author Rajiakodi et al. (2025).

We have used traditional machine learning techniques and deep learning like in the paper by Abeera et al. (2023) for sentiment analysis on the Malayalam and Tamil datasets. The methodology began with preprocessing, where we removed special characters and extra spaces, followed by normalizing the class labels to maintain consistency. Label encoding was applied to convert categorical class labels into numerical representations. Initially, we trained a FastText model, optimized for n-grams and dimensionality.

For the deep learning approach, we used pre-trained BERT model embeddings, combined with a Bi-LSTM with attention mechanism and a transformer encoding layer. The BERT tokenizer was utilized to tokenize the text and generate embeddings, which were then passed through the subsequent neural networks in the paper by Sheik et al. (2023). Later the dataset was splitted into 80% training and 20% testing.

3.1 Bi-LSTM Attention Architecture

For the Bi-LSTM with Attention model in the first layer consists of a pre-trained BERT model that generates contextual embeddings with an embedding dimension of 768 units. The next layer is a bi-directional LSTM with a hidden dimension of 256 units which capturing sequential dependencies from both forward and backward directions. An attention mechanism is applied to compute the weighted importance of LSTM outputs which allows the model to focus on the most relevant parts of the input sequence. A dropout layer with a rate of 0.3 is incorporated to prevent overfitting. The final fully connected layer maps the context vector to the number of unique classes in the dataset, using a softmax activation for classification.

3.2 Transformer Model Architecture

For the Transformer-Based Model like in Rajalakshmi et al. (2022) , the first layer consists of BERT embeddings as input, with an embedding dimension of 768 units. The next layer is a multi-head self-attention mechanism with 8 attention heads, allowing the model to capture contextual dependencies across the input sequence. A layer normalization step is applied to stabilize training. Following this, an additional Transformer encoder layer is used, which consists of a multi-head attention

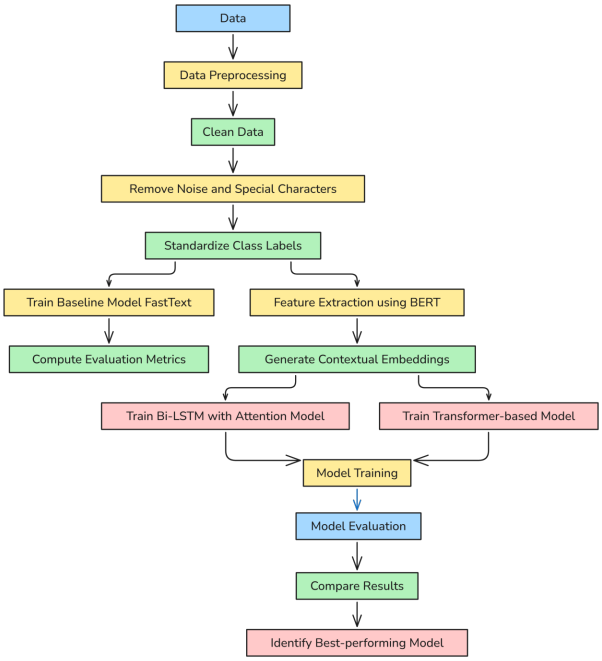


Figure 1: Proposed Methodology

mechanism and a position-wise feed-forward network with 4 times the hidden dimension (3072 units). The output is processed by two fully connected layers, where the first layer has 384 units with ReLU activation, and the second layer maps to the number of unique classes. A dropout layer with a rate of 0.3 is applied after each attention and feed-forward block.

3.3 FastText Architecture

We used FastText which was used in the paper by the Bojanowski et al. (2016), a word embedding and text classification library, to build a model for identifying abusive and non-abusive text in Tamil and Malayalam. The dataset was preprocessed by cleaning text, removing noise, and structuring it into labeled training and validation sets. FastText's supervised training was applied using bigram features, a 300-dimensional word vector, with a hierarchical softmax classifier to enhance classification accuracy. The model was trained over 50 epochs with an optimized learning rate and bucket size for better generalization. After training, validation was performed using precision, recall, and F1-score metrics. The final model was evaluated against a test dataset, and predictions were compared with ground truth labels to compute accuracy. The approach ensures an efficient and scalable solution for abusive language detection in low-resource languages.

4 Results

Table 1: Bi-LSTM with Attention Model Performance

Class	Precision	Recall	F1-Score	Support
Abusive	**0.55**	**0.80**	**0.65**	628
Non-Abusive	**0.61**	**0.33**	**0.42**	599
Accuracy			**0.57**	1227
Macro Avg	**0.58**	**0.56**	**0.54**	1227
Weighted Avg	**0.58**	**0.57**	**0.54**	1227

4.1 Hyperparameters

The Bi-LSTM with Attention model and the Transformer-based model were trained using the AdamW optimizer with a cross-entropy loss function, a learning rate of 2e-5, and for 15 epochs. The models were tested and evaluated based on accuracy, macro-precision, macro-recall, and macro-F1 score.

4.2 Performance Analysis

From Tables 1 and 2, which show the results of the models trained on a combination of Tamil and Malayalam datasets, we can see that the Trans-

Table 2: Transformer-Based Model Performance

Class	Precision	Recall	F1-Score	Support
Abusive	**0.59**	**0.80**	**0.68**	628
Non-Abusive	**0.67**	**0.43**	**0.52**	599
Accuracy			**0.62**	1227
Macro Avg	**0.63**	**0.61**	**0.60**	1227
Weighted Avg	**0.63**	**0.62**	**0.60**	1227

Table 3: Malayalam Dataset Evaluation Scores Using FastText

Class	Precision	Recall	F1-Score	Support
Abusive	**0.66**	**0.63**	**0.65**	323
Non-Abusive	**0.63**	**0.66**	**0.64**	305
Accuracy			**0.64**	628
Macro Avg	**0.65**	**0.65**	**0.64**	628
Weighted Avg	**0.65**	**0.64**	**0.64**	628

former model performed better than the Bi-LSTM with Attention model. The Transformer model achieved a macro F1-score of 0.60, while the Bi-LSTM model scored 0.54. Looking at each class, the Transformer model scored 0.68 for abusive and 0.52 for non-abusive, while the Bi-LSTM model got 0.65 for abusive and 0.42 for non-abusive. The table-3 and 4 represents the scores of the tamil and malaylam dataset in which the model achieved an macro f1 score of 0.73 for malayalam and the 0.64 score for the tamil dataset. From the tables we could infer that the fasttext model got better f1-scores when compared to other models because of the subword information capability. Additionally, fasttext's efficient word vectorization technique, it can able to captures semantic meaning in low-resource languages more effectively than models like Bi-LSTM or Transformers.

5 Conclusion

The urgent need to combat gender-based violence in digital spaces is underscored by the widespread

Table 4: Tamil Dataset Evaluation Scores Using FastText

Class	Precision	Recall	F1-Score	Support
Abusive	**0.74**	**0.75**	**0.74**	304
Non-Abusive	**0.73**	**0.72**	**0.73**	293
Accuracy			**0.74**	597
Macro Avg	**0.74**	**0.74**	**0.74**	597
Weighted Avg	**0.74**	**0.74**	**0.74**	597

abuse and harassment of women on Tamil and Malayalam social media platforms. The absence of effective systems to identify and reduce abusive content in regional languages contributes to this problem, which has its roots in social norms. The creation of sophisticated, language-specific NLP models is required due to the considerable difficulties that the linguistic complexity of Tamil and Malayalam, code-mixing, and cultural quirks present for current moderation systems.

By addressing this issue, we hope to encourage women to actively participate in online venues without fear of harassment, in addition to protecting them from targeted abuse. To create inclusive and context-aware solutions that guarantee safer and more equal digital environments, researchers, legislators, and technology developers must work together. This endeavor aims to promote equality, dignity, and respect in online conversation in addition to technology developments.

Our project is anonymously available at : https://tinyurl.com/3tnmvwpm

In conclusion, our test results show that FastText performs better than other models, with an accuracy of 0.64 on the Malayalam dataset and 0.74 on the Tamil dataset. FastText is still a popular option because of its capacity to handle words that are not in the lexicon and capture subword information, which is especially advantageous for morphologically rich Dravidian languages, even though it has limitations in memory consumption and semantic understanding. In order to enhance automatic moderation and guarantee safer online environments, these results emphasize the significance of language-specific NLP solutions designed for Dravidian languages.

6 Limitations

The Transformer Model's fundamental weakness was its inability to resolve the OOV issue. By resolving this issue, we increased the model's ability to handle unseen words, resulting in better generalization. The FastText model's fundamental weakness, however, was its use of large amounts of memory. The high memory requirement was due to its reliance on n-grams, potential misclassifications when words with similar n-grams are encountered, less nuanced semantic understanding than more complex models, and a linear classification model that may fail to capture complex relationships effectively in certain scenarios.

References

A. V. P. Abeera, S. Kumar, and K. P. Soman. 2023. Social media data analysis for malayalam youtube comments: Sentiment analysis and emotion detection using ml and dl models.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

B. R. Chakravarthi, R. Priyadharshini, S. Banerjee, M. B. Jagadeeshan, P. K. Kumaresan, R. Ponnusamy, S. Benhur, and J. P. McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

J. Mohan, S. Reddy Mekapati, P. B., J. L. G., and B. R. Chakravarthi. 2025. A multimodal approach for hate and offensive content detection in tamil: From corpus creation to model development.

B. Premjith, G. Jyothish, V. Sowmya, B. R. Chakravarthi, K. Nandhini, R. Natarajan, A. Murugappan, B. Bharathi, S. Rajiakodi, R. Ponnusamy, J. Mohan, and R. Mekapati. 2024. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@dravidianlangtech 2024.

R. Priyadharshini, B. R. Chakravarthi, S. Cn, T. Durairaj, M. Subramanian, K. Shanmugavadivel, S. U. Hegde, and P. Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022.

R. Priyadharshini, B. R. Chakravarthi, S. C. Navaneethakrishnan, M. Subramanian, K. Shanmugavadivel, B. Premjith, A. Murugappan, S. P. Karnati, Rishith, J. Chandu, and P. K. Kumaresan. 2023. Findings of the shared task on abusive comment detection in tamil and telugu.

R. Rajalakshmi, S. Selvaraj, R. Mattins, P. Vasudevan, and M. Kumar. 2022. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming.

S. Rajiakodi, B. R. Chakravarthi, S. Muthusamy Chinnan, R. Priyadharshini, J. Rajameenakshi, K. Pannerselvam, R. Ponnusamy, B. Sivagnanam, P. Buitelaar, K. Bhavanimeena, V. Jananayagam, and K. Ponnusamy. 2025. Findings of the shared task on abusive tamil and malayalam text targeting women on social media: Dravidianlangtech@naacl 2025.

R. Sheik, R. Balanathan, and S. Nirmala. 2023. Mitigating abusive comment detection in tamil text: A data augmentation approach with transformer model.

M. Subramanian, R. Chinnasamy, K. Shanmugavadivel, N. Subbarayan, A. Ganesan, D. Ravi, V. Palanikumar, and B. R. Chakravarthi. 2022. On finetuning adapter-based transformer models for classifying abusive social media tamil comments. SSRN.

A. Vetagiri, G. Kalita, E. Halder, C. Taparia, P. Pakray, and R. Manna. 2024. Breaking the silence: Detecting and mitigating gendered abuse in hindi, tamil, and indian english online spaces. arXiv preprint arXiv:2404.02013.

LinguAIsts@DravidianLangTech 2025: Misogyny Meme Detection using multimodel Approach

ARTHI R[1], **Pavithra J** [2], **G Manikandan** [3], **Lekhashree A**[4]
Dhanyashree G[5], **Bommineni Sahitya**[6], **Arivuchudar K**[7], **Kalpana K**[8]
R.M.K. Engineering College, Tiruvallur, Tamilnadu, India
{arth22004, pavi22039, mgk, lekh22026}.ad@rmkec.ac.in
{dhan22012, bomm22009, ariv22002, kalp22020}.ad@rmkec.ac.in

Abstract

Memes often disseminate misogynistic material, which nurtures gender discrimination and stereotyping. While it is an effective tool of communication, social media has also provided a fertile ground for online abuse. This vital issue in the multilingual and multimodal setting is tackled by the Misogyny Meme Detection Shared Task. Our method employs advanced NLP techniques and machine learning models to classify memes in Malayalam and Tamil, two low-resource languages. Preprocessing of text includes tokenization, lemmatization, and stop word removal. Features are then extracted using TF-IDF. With the best achievable hyperparameters, along with the SVM model, our system provided very promising outcomes and ranked 9th among the systems competing in the Tamil task with a 0.71259 F1-score, and ranked 15th with an F1-score of 0.68186 in the Malayalam taks. With this research work, it would be established how important AI-based solutions are toward stopping online harassment and developing secure online spaces.

1 Introduction

Social media has enabled international artistic exchange but is also a platform for the circulation of harmful content, especially gender-based abuse. There is growing concern about the proliferation of misogynistic memes, which are a group of images and text that support anti-woman discourses Ponnusamy et al., 2024. Since they reinforce negative gender norms and stereotypes, their identification becomes crucial in widespread terms and Chakravarthi, 2021.

Earlier research dealt with all dimensions of on- line hate speech detection. Transformer models such as BERT has made tremendous progress in language understanding (Devlin et al., 2019). It is evident that deep learning techniques have been used to identify offensive language . The detection of hate speech in multimodal content has also been studied, emphasizing how such models need to be able to process both text and images Gomez et al., 2020. Further, multilingual NLP approaches, like MuRIL, have advanced abusive content detection in Indian lan- guages (Khanuja et al., 2021). But for misogynistic memes, there is a specific approach that is needed, one that combines both linguistic and visual fea- tures Suryawanshi et al., 2021.

Figure 1: Sample for misogynistic meme in Tamil content.

The situation is particularly challenging for low- resource languages like Tamil and Malayalam, where content moderation technologies are rela- tively weak compared to high-resource languages (Thavareesan et al.,2019). Most languages glob- ally have inadequate datasets and tools necessary for filtering harmful content, which facilitates the propagations of misogynistic material unabatedly (Bishop, 2014). The issue is multilingual and mul- timodal; therefore, automated detection systems need to be built in order to counter gender-based abuse effectively.

Therefore, DravidianLangTech@NAACL 2025 introduces this Shared Task on Misogyny Meme Detection, focusing on low-resource languages to determine the distinction between misogynistic and non-misogynistic memes. This initiative not only addresses an urgent social issue but also fills a gap in research by developing AI-based systems for ethical content moderation in underrepresented lan- guage communities

We present an approach that relies on sophis- ticated Natural Language Processing (NLP) tech- niques to

Figure 2: Sample for non-misogynistic meme in Tamil con- tent.

analyze the textual aspect of memes. To- kenization splits the text into manageable units, lemmatization reduces words to their root forms while maintaining meaning, and Term Frequency- Inverse Document Frequency (TF-IDF) points out the most important terms in the dataset. For classifi- cation, we rely on Support Vector Machines (SVM), a machine learning algorithm famous for its ef- ficiency in high-dimensional data environments (Suryawanshi et al., 2021). Our work, through the incorporation of low- resource languages and multimodal analysis, offers a scalable and inclusive solution to gender-based abuse across various online platforms. Our study, titled Towards Protecting Marginalized Communi- ties: Mitigating Inferences from Large Language Models, aims to establish a robust system for de- tecting sexism in online content while laying the foundation for future applications in multilingual and multimodal AI research.

2 Related Work

Training a misogyny detectory feature on memes has had a lot of traction lately, especially with re- cent studies done on low resource language like Tamil and Malayalam, but we just scratched the surface. Datasets such as HASOC (Kumar et al., 2021) and Mandl et al. (2020) are oriented towards hate speech in Hindi and Bengali while the Hate- ful Memes Challenge (Kiela et al., 2020) aims to multimodal hate speech detection in English, trans- ferring these approaches to detect gender-based hate in multilingual settings is still a challenge. Some noteworthy examples include research such as Fersini et al. (2018)—while there is a multi-level sexism detection, English and Spanish based, such multilingual multimodal datasets targeting misogy- nistic memes in languages like Tamil and Malay- alam remain insufficient. Ponnusamy et al., 2024 does an excellent job at bridging this gap by pro- viding an annotated dataset for misogyny detection in Tamil and Malayalam memes. Overall, njihova dataset and framework provide a necessary step in the right direction for being able to witnes and identify sexist abuse in local social media mate- rial. We extend their work by introducing a SVM classifiers with TF-IDF feature extraction and text pre-processing for challenging low-resource languages. With these techniques, we hope to have a model that does better than others in the task of misogynistic detection in memes, thus vastly im proving the study of gender bias in multilingual platforms.

3 System Description

In this section, we give a detailed description of the dataset and offer more information about the experiments that were carried out for the study. In Figure 3, the system architecture for misogynistic and non-misogynistic meme classification utilising machine learning (ML) approaches, including Grid- SearchCV for hyperparameter tuning, is depicted. The whole classification process flow is depicted in the diagram, which highlights the crucial steps in identifying misogynistic content in low-resource 2 languages like Tamil and Malayalam.

3.1 Dataset

The dataset used in this research consists of text memes of Tamil and Malayalam. The dataset for all languages is split into train, validate and test parts. The dataset is annotated into 2 classes: Misogy- nistic and Non-Misogynistic.Table 1 gives distribu- tion details over the dataset and number of samples available in each subset for both languages.

Language	Train	Validate	Test
Tamil	1137	285	357
Tamil	641	161	201

Table 1: Dataset distribution for Tamil and Malayalam.

The datasets were adapted and modified from publicly accessible datasets originally published as part the DravidianLangTech@NAACL program to suit the specific context of this study. This program focuses on building language technology resources and tools for low-resource Dravidian languages.

4 Methodology

In general, there are four processes involved in Identification of Misogynistic and Non- Misogynistic memes in Tamil and Malayalam such as EDA, Preprocessing, Modelling and Evaluation. In EDA, we explored the dataset and analysed ob- servation of meme linguistics. Pre-processing To- kenisation and lemmatisation

were performed to clean and standardise the raw text. During the Modelling phase, Support Vector Machines (SVM) were used to classify the memes. In the Evaluation phase, we used metrics (Tuning phase) such as accuracy, F1 score, precision, and recall to evaluate the performance of the train model. F1 score of a balanced classification was the measure of per- formance. This comprehensive strategy ensured the accurate classification of memes while consid- ering the subtleties of the Tamil and Malayalam languages.

4.1 Exploratory Data Analysis

Through exploratory data analysis (EDA) it is found that there are some important linguistic patterns present in Misogynistic and Non-Misogynistic memes in Tamil and Malayalam. The common occurence of word pari/triples were determinated with N-gram Analysis while common associated words to sexist material were identified using Term Frequency Analysis. These attempts made informed additions to the set of features for training the models, such as highlighting abusive phrases and identifying certain linguistic patterns that significantly improved the meme classification accuracy.

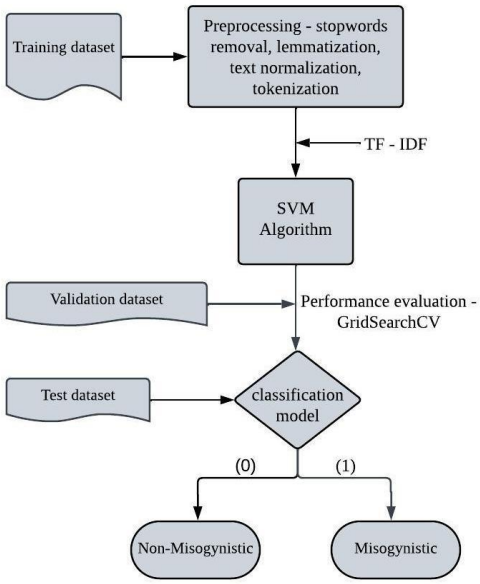


Figure 3: System Architecture for classify misogyny meme using ML models.

4.2 Preprocessing

We simplified the language by removing punctu- ation, converting all characters to lowercase, and removing numeric characters. In order to retain meaning when a variety of words were included, lemmatisation was used to reduce words to their lowest forms. The revised text was subsequently converted into numerical features leveraging TF- IDF vectorisation, using bigrams and unigrams to encapsulate both isolated and contextually corre- lated word associations.

4.3 Machine Learning Model

For the classfication of Misogynistic and Non-Misogynistic memes in Tamil and Malayalam we used Support Vector Machines(SVM) as the core machine learning based model. SVM was chosen for being effective with textual data and for its power to deal with the intricacies of classifying memes. The model was selected due to its capability of performing binary classification which was a task given to identify a meme as misogynistic or non-misogynistic.

- **Support Vector Machine (SVM)** The best hyperplane for separating the Misogynistic and Non-Misogynistic memes in the high- dimensional TF-IDF vector/matrix was found by using a linear kernel in the SVM model. This was indicative of strong text-based data handling, as well as generalization. Accura- cies were found to be 66fying Tamil memes and Malayalam memes re- spectively indicative of the difficulty faced in classifying memes in low resource languages.

4.4 Model Evaluation

Important metrics (accuracy, F1 score, precision, recall) were used for the assessing the model for balanced classification of Misogynistic and Non- Misogynistic memes. The accuracy for Tamil memes was 78resents a good performance, but reveals that there is work to do in translating between precision and recall. For Malayalam memes, the same model's F1 score was 71were somewhat better and showed that the model had handled the characteristics of the Malayalam language better than others. Cross-validation was used to ensure robustness, while the F1 score was prioritized to maintain the balance between preci- sion and recall.

5 Results

We evaluated our model effectiveness at Misogy- nistic and Non-Misogynistic meme detection us- ing the macro-average f1-score as our performance metric. Since there is a class imbalance in our data, we compute the macro F1-score: it calculates the F1-score for each class (Misogynistic and Non Misogynistic) and takes the mean of these scores to not let the imbalance influence the evaluation and to treat both classes equally. Using this method gives a more even gauge of performance across categories.

These results indicate that the model was able to distinguish between the two categories successfully, being able to detect Misogynistic memes with high recall but accuracy in Non-Misogynistic recognition.

5.1 AUC and ROC Curve

The ROC curve entails plotting the True Positive Rate (TPR, alternatively known as recall) against the False Positive Rate (FPR) at varying thresholds for classification. This approach is a visual way of examining

Labels	Accuracy	F1-Score	Precision	Recall
Non-Misogynistic	0.78	0.87	0.80	0.95
Misogynistic	0.78	0.44	0.71	0.32
Macro average	0.78	0.66	0.75	0.64
Weighted average	0.78	0.76	0.78	0.79

Table 2: Tamil memes misogynistic content classification report

Labels	Accuracy	F1-Score	Precision	Recall
Non-Misogynistic	0.73	0.78	0.77	0.78
Misogynistic	0.73	0.65	0.66	0.63
Macro average	0.78	0.66	0.75	0.64
Weighted average	0.73	0.72	0.72	0.72

Table 3: Tamil memes misogynistic content classification report

the trade-offs between sensitivity and specificity for a model. A model AUC (goods under the curve) of 0.88 indicates a pretty good performance, meaning it is able to discriminate between Misogynistic memes and Non-Misogynistic memes very efficiently. The closer the AUC is to 1, the model has a higher discriminative power, which means it accurately classifies positive and negative instances. A higher AUC denotes better model performance, while an AUC of 0.5 simply means a random guess. By looking at the ROC curve, we will gain more knowledge about how different thresholds could affect classification performance, thereby ensuring better threshold selection. This curve will allow one to balance between detecting misogynistic memes and lowering false positives.

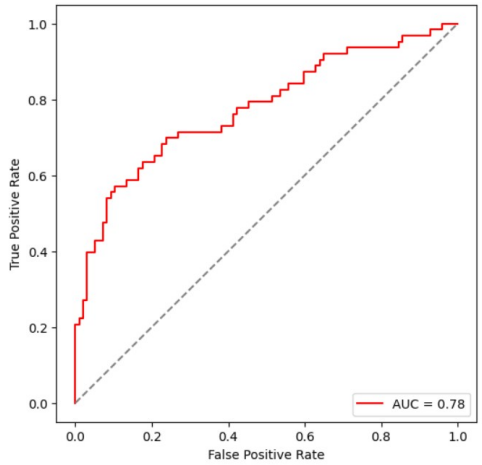


Figure 4: AUC and ROC curve for Malayalam memes, indicating the model's classification performance.

5.2 Confusion Matrix

Our misogyny meme classification model is tested on the basis of the Confusion Matrix, which shows the count of True Positives (TP), True Neg- atives (TN), False Positives (FP), and False Nega- tives (FN). These are the counts needed to calcu- late significant performance metrics such as ac- curacy, precision, recall, and F1 score. These metrics help in identifying how effectively the model distinguishes between misogynistic and non- misogynistic memes, giving an accurate view of its classification performance.

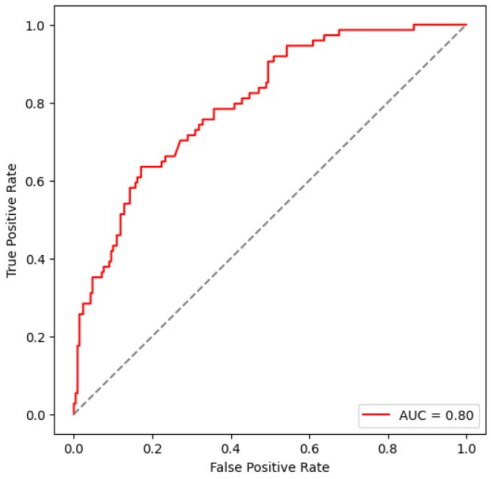


Figure 5: Measures the model's performance in distinguishing misogynistic content in Tamil memes.

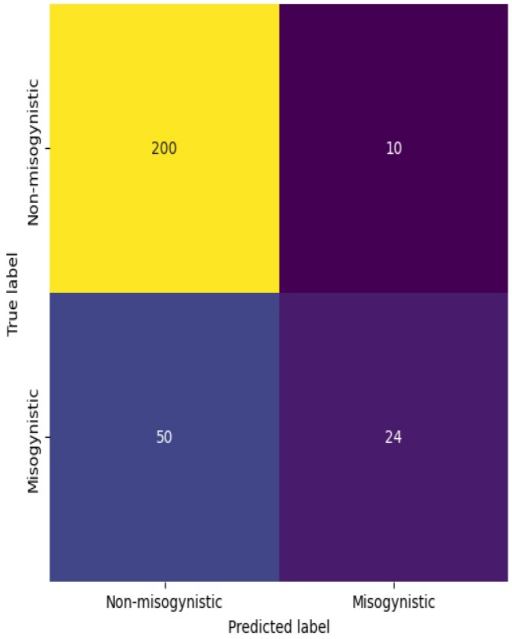


Figure 6: Displays correctly and incorrectly classified misogynistic and non-misogynistic memes in Tamil.

6 Future work

We plan to broaden the scope of our misogyny meme-detecting system to improve its performance and flexibility with different languages and differ- ent types of memes in future work. The idea can be applied to multimodal analysis, where text- and image-based content within a meme for classifica- tion are taken into account. This can be done by us- ing advanced transformer-based models such as Vi- sualBERT or CLIP, which is quite

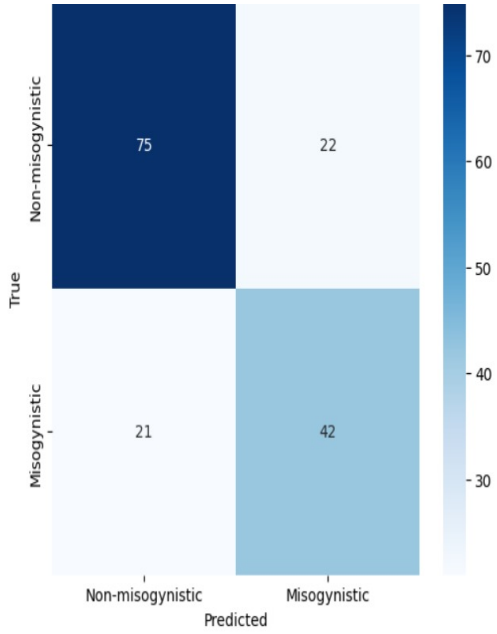


Figure 7: Displays correctly and incorrectly classified misogynistic and non-misogynistic memes in Tamil.

promising in han- dling multimodal data. We would like to improve the model by using more diverse datasets and fine- tune the model with domain-specific datasets to bet- ter capture regional flavor in languages like Tamil and Malayalam. An another way for the improve- ment is by using active learning approaches that reduce manual effort in interpreting new memes so that the system can continue to improve as more data is present. Finally, investigating real-time meme detection could open up possibilities for the practical application of the system, For example, in social media moderation tools that could help fight gender-based abuse online more effectively.

7 Conclusion

This study addresses the very important issue of detection of misogynistic content in memes, especially in two low resource languages, Malayalam and Tamil which are generally neglected by the present-day content moderation systems. Using a combination of SVM classification, text preparation and data EDA the method successfully detects dangerous language hidden in the meme content. It shows the problems and intricacies of addressing abusive content in linguistically and culturally diverse environments. This work aids in the identification of gender-based discrimination and emphasizes the demand for continued re- search and development of more precise and context aware meme categorization models. Additional datasets, language-related components, and multi-modal analysis.

8 Limitations

While advances have occurred, several limita- tions persist. Low-resource systems face chal- lenges when adapting themselves to multiple low- resource languages due to the lack of large anno- tated datasets and linguistic variance in classifica- tion accuracy. Hence, high computational capabil- ity and resource training are involved in handling multimodal data, raising concerns over scalability. Biases in training data will often produce inconsis- tency in the detection of misogyny in differing cul- tural contexts, restricting the system's generaliza- tion, even when trained on heterogeneous datasets. However, aiding the process with adaptive learning techniques will still require heavy human interven- tion to double-check and amend model predictions, which will be quite tedious. Real-time detection re- sults in latency, which inhibits the efficient process- ing of large amounts of data. Another challenge is that the content is updated frequently with new words, symbols, and hidden connotations, which also need to be updated frequently to maintain their accuracy. For in-text citation, add Wu et al. (2006) after "huge ramifications". Include a few sentences about the possible ramifications of wrongful classi- fication. Wrongly characterizing presumably gen- erous content can jeopardize its trustworthiness. Erroneously flagging harmful content as benign puts an additional dent in the trustworthiness of the system. Any kind of building of a content modera- tion system needs to implicate ethics and privacy directly. It must weigh censorship against free ex- pression properly.

Acknowledgment

We thank DravidianLangTech-2025 at NAACL 2025 shared task organizers for providing data sets and guidance.https://sites.google.com/view/dravidianlangtech-2025/shared-tasks-2025

References

Ponnusamy, Rahul and Kathiravan Pannersel- vam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, An- shid K.A, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From Laughter to Inequality: Annotated Dataset for Misogyny Detection in Tamil and Malayalam Memes. *In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evalua- tion (LREC-COLING 2024)*, pages 7480-7488.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. *Multi- model Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Textf.*, pages 32-41.

Jonathan Bishop. 2014. Dealing with internet trolling in political online communities: To- wards the this is why we can't have nice things scale *International Journal of E-Politics (IJEP), 5(4):1-20.*

Sajeetha Thavereesan and Sinnathamby Mahe- san. 2019.Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation.*In 2019 14th Confer- ence on Industrial and Information Systems (ICIIS)*, pages 320-325.

Thomas Davidson, Dana Warmsley, M. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *In Proceedings of the International AAAI Con- ference on Web and Social Media (ICWSM).*

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).*

Sajeetha Thavareesan and Sinnathamby Mahasen. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th Inter- national Conference on Industrial and Informa- tion Systems(ICIIS),*, pages 478– 482.

Suryawanshi2021 Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. Association for Computational Linguistics.*

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Ra- jiv Teja Nagipogu, Shachi Dave.In *Multilingual representations for Indian languages. arXiv preprint arXiv:2103.10730.*

B Bharathi and A Agnusimmaculate Silvia. 2021. SSNCSE NLP@DravidianLangTech- EACL2021: Offensive language identification 7 on multilingual code-mixing text. In*Proceed- ings of the First Workshop on Speech and Lan- guage Technologies for Dravidian Languages, pages 313–318, Kyiv.*

E. S. Smitha, S. Sendhilkumar, and G. S. Maha- laksmi. 2018. Meme classification using textual and visual features. In *Computational Vision and Bio Inspired Computing, Cham. Springer International Publishing*, pages 1015–1031.

Suryawanshi2021 Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. A Sentiment Analysis Dataset for code-Mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under- resourced languages (SLTU) and Collabora- tion and Computing for Under-Resourced Lan- guages (CCURL),* pages 177–184, Marseille, France.

Chakravarthi, Bharathi Raja, Ponnnusamy, Rahul and Rajiakodi, Saranya and Muthusamy, Sivagnanam, Bhuvaneswari and Kizhakkeparambil, Anshid. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Preceedings of the Fifth Workshop on Speech, Vision and Language Technologies for Dravidian Languages* Association for Computational Linguistics, 2025.

CUET_Agile@DravidianLangTech 2025: Fine-tuning Transformers for Detecting Abusive Text Targeting Women from Tamil and Malayalam Texts

Tareque Md Hanif[1] and **Md Rashadur Rahman**[2]
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
[1]tarequemd.hanif@yahoo.com
[2]rashadur@cuet.ac.bd

Abstract

As social media has grown, so has online abuse, with women often facing harmful online behavior. This discourages their free participation and expression online. This paper outlines the approach adopted by our team for detecting abusive comments in Tamil and Malayalam. The task focuses on classifying whether a given comment contains abusive language towards women. We experimented with transformer-based models by fine-tuning Tamil-BERT for Tamil and Malayalam-BERT for Malayalam. Additionally, we fine-tuned IndicBERT v2 on both Tamil and Malayalam datasets. To evaluate the effect of pre-processing, we also conducted experiments using non-preprocessed text. Results demonstrate that IndicBERT v2 outperformed the language-specific BERT models in both languages. Pre-processing the data showed mixed results, with a slight improvement in the Tamil dataset but no significant benefit for the Malayalam dataset. Our approach secured first place in Tamil with a macro F_1-score of 0.7883 and second place in Malayalam with a macro F_1-score of 0.7234. The implementation details of the task will be found in the GitHub repository.[1]

1 Introduction

Over the past decade, the exponential growth of user-generated content on social media has unfortunately led to increased abusive behavior online. This includes cyberbullying, hate speech, and offensive language, often targeting various classes of people including women. These actions can lead to real-world violence and push women to the sidelines, making them feel excluded and undervalued both online and in everyday life (Kaur et al., 2021). A study in 51 countries found that 38% of women have faced online harassment. Only 25% of them reported it, and 90% reduced their online activity (Hashmi et al., 2024).

Tamil is among the oldest languages in the world, spoken by over 65 million people globally (Ramakrishnan et al., 2007). Malayalam, the official language of Kerala, has more than 37 million speakers worldwide (Rojan et al., 2020). Both Tamil and Malayalam have many dialects, making it challenging to develop NLP systems for these languages.

Developing an intelligent abuse detection model is challenging in resource-constrained languages like Tamil and Malayalam. Therefore, a shared task was organized to encourage the development of effective abuse detection models for these languages.

This shared task (Rajiakodi et al., 2025) aims to detect abusive comments targeting women in Tamil and Malayalam, sourced from YouTube comments. The dataset contains text in both languages, with each comment classified as either 'Abusive' or 'Non-Abusive'. The task focuses on identifying explicit abuse, implicit bias, stereotypes, and coded language directed at women on social media.

We fine-tuned transformer-based models for text classification. Specifically, we used Tamil-BERT (Joshi, 2022) for Tamil comments and Malayalam-BERT (Joshi, 2022) for Malayalam comments. Additionally, we fine-tuned IndicBERT v2 (Doddapaneni et al., 2023) on both Tamil and Malayalam datasets. We also experimented with training models on non-preprocessed text to analyze the impact of preprocessing.

The rest of the paper is organized into 6 sections. Section 2 reviews related work in Natural Language Processing, focusing on misogynistic text detection in Tamil, Malayalam, and other languages. Section 3 describes the dataset provided by the shared task organizers. Section 4 provides a detailed explanation of the proposed methodology and the models implemented. Section 5 presents the results and key observations. Finally, Section 6 concludes the paper.

[1]https://github.com/tmdh/DravidianLangTech-NAACL-2025-ATTW

2 Related Works

Recent advances in NLP have increased interest in detecting different types of hate speech, leading to many new and creative methods in this field. Offensive language detection in Tamil and Malayalam has been studied in previous research (Ponnusamy et al., 2024), but to the best of our knowledge, this is the first shared task that specifically focuses on detecting abusive texts targeting women from Dravidian texts. There have been previous shared tasks on languages other than Dravidian for misogynistic text detection, such as the Arabic Misogyny Identification (ArMI) task (Mulki and Ghanem, 2022) and the GermEval2024 shared task, GerMS-Detect (Gross et al., 2024). The ArMI task combined two subtasks: a binary classification for detecting misogynistic language and a multi-class classification for identifying seven misogynistic behaviors in 9,833 Arabic/dialectal tweets. GerMS-Detect focused on detecting sexism and misogyny in German language online news comment.

In terms of Dravidian languages, Chakravarthi et al., 2023 proposed a fusion model of MPNet (Song et al., 2020) and CNN for offensive language identification in code-mixed Tamil, Malayalam, and Kannada social media comments, achieving superior results over classical ML and transformer-based baselines.

Sreelakshmi et al., 2024 explored offensive language detection in code-mixed Tamil-English, Malayalam-English and Kannada-English using multilingual transformer embeddings with Support Vector Machine classifiers, identifying MuRIL (Khanuja et al., 2021) as the most effective model across various datasets.

The study by Vasantharajan and Thayasivam, 2021 explores offensive language detection in Tamil code-mixed YouTube comments, proposing selective translation and transliteration techniques to enhance transformer models like BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). Their findings highlight ULMFiT (Howard and Ruder, 2018) and mBERT-BiLSTM as the most effective models for offensive language detection.

Prior work in abuse detection has primarily focused on English, lacking substantial datasets for Indic languages. Gupta et al., 2022 proposed MACD to address this gap by introducing a large-scale multilingual abuse detection dataset and AbuseXLMR model for Indic languages.

Class	Train	Dev	Test
Abusive	1366	278	305
Non-Abusive	1424	320	293
Total	2790	598	598

Table 1: Class-wise distribution of Tamil Dataset

Class	Train	Dev	Test
Abusive	1531	303	323
Non-Abusive	1402	326	306
Total	2933	629	629

Table 2: Class-wise distribution of Malayalam Dataset

3 Dataset Description

The organizers of the Abusive Text Targeting Women Detection shared task provided two datasets (Priyadharshini et al., 2023, Priyadharshini et al., 2022) where one consists of Tamil texts while the other consists of Malayalam texts. Each of the texts is annotated with one of the the classes: Abusive and Non-Abusive. Table 1 displays the class-wise data distribution for the Tamil dataset, while Table 2 shows the same for the Malayalam dataset.

To provide better insights, we conducted a more in-depth analysis of the training set. Table 3 presents the detailed statistics of the training data.

4 Methodology

This work employed two transformer-based models on each language's dataset, both preprocessed and non-preprocessed. Firstly, we removed unwanted characters (i.e., numbers, extra spaces, and URLs) from the texts in both the Tamil and Malayalam datasets to create two preprocessed datasets.

4.1 Transformer Models

Recent advancements in NLP have shown that transformer-based models perform better than other approaches for text classification across different languages. In this work, we fine-tuned Tamil-

Statistics	Abusive	Non-Abusive
Total words	21166	19091
Unique words	9541	8672
Max. length (words)	48	48
Avg. words (per text)	15.5	13.4

Table 3: Detailed statistics of each class in the training set

Approach	Selected Epoch	Accuracy	Precision	Recall	F_1-score
Tamil-BERT (Non-preprocessed)	4	0.7793	0.7800	0.7786	0.7788
Tamil-BERT (Preprocessed)	2	0.7843	0.7861	0.7850	0.7842
IndicBERT v2 (Non-preprocessed)	3	0.7876	0.7945	0.7860	0.7857
IndicBERT v2 (Preprocessed)	2	**0.7893**	**0.7923**	**0.7882**	**0.7883**

Table 4: Performance comparison of various models on the test set of Tamil dataset

Approach	Selected Epoch	Accuracy	Precision	Recall	F_1-score
Malayalam-BERT (Non-preprocessed)	2	0.6630	0.7136	0.6692	0.6467
Malayalam-BERT (Preprocessed)	5	0.6439	0.6440	0.6426	0.6424
IndicBERT v2 (Non-preprocessed)	2	**0.7234**	**0.7238**	**0.7239**	**0.7234**
IndicBERT v2 (Preprocessed)	4	0.7122	0.7133	0.7130	0.7122

Table 5: Performance comparison of various models on the test set of Malayalam dataset

BERT[2] and Malayalam-BERT[3] on Tamil and Malayalam texts, respectively, using both preprocessed and non-preprocessed datasets. We also used IndicBERT v2[4], a multilingual model, to handle both languages.

Our classifier is based on a transformer model with a linear classification head. The architecture consists of a pre-trained BERT model followed by a fully connected layer that maps the hidden state of the [CLS] token to a two-class output. The model was trained using PyTorch Lightning (Falcon and team, 2024), which simplified the training and evaluation process. We optimized the models using the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of $5e-5$. The training process ran for up to five epochs, and we selected the best-performing epoch based on the highest F_1-score on the validation set.

For training, we used a batch size of 32 and applied cross-entropy loss. The training process logged F_1-score on the validation set at each epoch. Model checkpoints were saved after each epoch, and the model with the highest F_1 was used for evaluation.

Table 6 summarizes the hyperparameters used across all models. The selected epochs for each approach are shown in Table 4 and Table 5 for Tamil and Malayalam, respectively.

[2] https://huggingface.co/l3cube-pune/tamil-bert

[3] https://huggingface.co/l3cube-pune/malayalam-bert

[4] https://huggingface.co/ai4bharat/IndicBERTv2-MLM-only

Hyperparameters	Values
Learning Rate	$5e-5$
Batch Size	32
Max Epochs	5
Weight Decay	0.01

Table 6: Hyperparameters used across models

4.1.1 Tamil-BERT and Malayalam-BERT

Tamil-BERT and Malayalam-BERT are monolingual BERT models fine-tuned from the multilingual MuRIL model for the Tamil and Malayalam languages, respectively. They are trained on large monolingual corpora. These models aim to enhance performance on downstream NLP tasks for these low-resource Indian languages. (Joshi, 2022)

4.2 IndicBERT v2

IndicBERT v2 is a state-of-the-art multilingual language model designed specifically for Indic languages. It supports all 24 languages covered in the IndicCorp v2 dataset. The dataset includes 20.9 billion tokens from 24 languages, including Indian English. This model is a significant step forward in building robust NLU capabilities for diverse Indic languages. (Doddapaneni et al., 2023)

5 Results

Table 4 and 5 reports the performance comparison of the different approaches on the Tamil dataset and Malayalam dataset, respectively. The effectiveness of the models is determined based on the macro F_1-score.

For the Tamil dataset, IndicBERT v2 fine-tuned on the preprocessed dataset achieved the high-

est F_1-score of 0.7883, followed by IndicBERT v2 on the non-preprocessed dataset with an F_1-score of 0.7857. For the Malayalam dataset, IndicBERT v2 employed on the non-preprocessed dataset achieved the best performance with an F_1-score of 0.7234, while IndicBERT v2 on the pre-processed dataset also performed well with an F_1-score of 0.7122. It is indicating that pre-processing did not improve the performance on the Malayalam dataset. For both Tamil and Malayalam, Tamil-BERT and Malayalam-BERT did not perform well on the task, while IndicBERT v2 achieved strong performance in both languages.

6 Conclusion

This paper investigated two language specific transformer models and one multilingual language model to detect abuse targeted towards women from preprocessed and non-preprocessed Tamil and Malayalam texts. Among all approaches, the highest macro F_1-score 0.7883 for Tamil texts is obtained by IndicBERT v2 fine-tuned with preprocessed Tamil texts. For Malayalam texts, the highest macro F_1-score 0.7234 is gained by finetuning IndicBERT v2 with non-preprocessed Malayalam texts. Looking ahead, we plan to explore ensemble methods and other advanced transformer-based models including MuRIL and XLM-R.

Limitations

While our model demonstrated strong performance in identifying abusive text directed at women in Tamil and Malayalam, it is important to recognize several limitations. These include the limited availability of diverse and high-quality annotated datasets for Dravidian languages, which restricts the model's ability to generalize across various dialects. Furthermore, the linguistic intricacies of Tamil and Malayalam can affect the model's effectiveness, especially in detecting implicit or subtly coded abusive language. Another challenge is the scalability of transformer-based models when handling longer texts, as they are mainly designed and optimized for shorter sequences. Moreover, fine-tuning these models demands significant GPU resources, which could restrict access for researchers with limited computational capabilities. Additionally, we did not perform an extensive hyperparameter search for critical parameters like learning rate and weight decay, which might otherwise enhance the model's performance.

References

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

William Falcon and The PyTorch Lightning team. 2024. Pytorch lightning.

Stephanie Gross, Johann Petrak, Louisa Venhoff, and Brigitte Krenn. 2024. GermEval2024 shared task: GerMS-detect – sexism detection in German online news fora. In *Proceedings of GermEval 2024 Task 1 GerMS-Detect Workshop on Sexism Detection in German Online News Fora (GerMS-Detect 2024)*, pages 1–9, Vienna, Austria. Association for Computational Lingustics.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, hastagiri prakash vanchinathan, and Animesh Mukherjee. 2022. Multilingual abusive comment detection at scale for indic languages. In *Advances in Neural Information Processing Systems*, volume 35, pages 26176–26191. Curran Associates, Inc.

Ehtesham Hashmi, Muhammad Mudassar Yamin, Shariq Imran, Sule Yildirim Yayilgan, and Mohib Ullah. 2024. Enhancing misogyny detection in bilingual texts using fasttext and explainable ai. In *2024 International Conference on Engineering & Computing Technologies (ICECT)*, pages 1–6.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Simrat Kaur, Sarbjeet Singh, and Sakshi Kaushal. 2021. Abusive content detection in online user-generated data: A survey. *Procedia Computer Science*, 189:274–281. AI in Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Hala Mulki and Bilal Ghanem. 2022. Working notes of the workshop arabic misogyny identification (armi-2021). In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 7–8, New York, NY, USA. Association for Computing Machinery.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

A.G. Ramakrishnan, Lakshmish Kaushik, and Laxmi Narayana. 2007. Natural language processing for tamil tts.

Annlin Rojan, Edwin Alias, Georgy M. Rajan, Jithin Mathew, and Dhanya Sudarsan. 2020. Natural language processing based text imputation for malayalam corpora. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 161–165.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Preprint*, arXiv:2004.09297.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Charangan Vasantharajan and Uthayasanker Thayasivam. 2021. Towards offensive language identification for tamil code-mixed youtube comments and posts. *SN Computer Science*, 3(1):94.

Necto@DravidianLangTech 2025: Fine-tuning Multilingual MiniLM for Text Classification in Dravidian Languages

Livin Nector Dhasan
IIT Madras / BS Degree
livin@study.iitm.ac.in

Abstract

This paper explores the application of a fine-tuned Multilingual MiniLM model for various binary text classification tasks, including AI-generated product review detection, abusive language targeting woman detection, and fake news detection in the Dravidian languages Tamil and Malayalam. This work was done as part of submissions to shared tasks organized by DravidianLangTech@NAACL 2025. The model was fine-tuned using both Tamil and Malayalam datasets, and its performance was evaluated across different tasks using macro F1-score. The results indicate that this model produces performance very close to the best F1 score reported by other teams. An investigation is conducted on the AI-generated product review dataset and the findings are reported.

1 Introduction

The advancement of natural language processing (NLP) models has significantly improved text classification capabilities. However, Dravidian languages, such as Tamil and Malayalam, remain underrepresented in NLP research. BERT-based models like mBERT, XLM-Roberta, and IndicBERT, have been demonstrating significant results in different classification tasks in the context of fake news detection(Luo and Wang, 2023; Tabassum et al., 2024) and abusive content detection(Hegde et al., 2023). To reduce the computational costs in the fine-tuning and the inference, smaller models of the BERT family with a lesser number of parameters such as DistilBERT(Sanh et al., 2020), MobileBERT(Sun et al., 2020) and TinyBERT(Jiao et al., 2019) are pre-trained using different knowledge distillation methods with a larger BERT based model as the Teacher and then fine-tuned for downstream tasks. These distilled models show near-comparable performance with fewer number of parameters.

The model used in this study, Multilingual MiniLM, was pre-trained using the Deep Self-Attention Distillation method by distilling the XLM-Roberta model (Wang et al., 2020). It outperforms similar distilled models such as DistilBERT, TinyBERT, and MobileBERT on various benchmarks. The multilingual nature of MiniLM made it a suitable choice for fine-tuning with Tamil and Malayalam data. This paper investigates the effectiveness of the Multilingual MiniLM model for diverse text classification tasks in these languages. By fine-tuning the model, specific challenges such as detecting AI-generated content, abusive language targeting women, and fake news are addressed.

2 Task Description

2.1 AI-Generated Product Review Detection (ai-gen)

This task addresses the growing concern of AI-generated product reviews, particularly in Tamil and Malayalam. As AI tools for content generation become more sophisticated, distinguishing between human-written and AI-generated reviews has become essential to ensure authenticity and reliability in consumer decision-making. The dataset poses it as a binary sentence classification problem, classifying the given product review text as "HUMAN" or "AI" containing data splits for both Tamil and Malayalam languages (Premjith et al., 2025). The data set does not include a development split for both languages, thus a development set is created from the training set using a stratified 80-20 train-dev split, which is used for fine-tuning.

2.2 Abusive Text Targeting Women on Social Media (abusive-woman)

This task focuses on classifying social media texts, particularly comments on YouTube, that are directed at women in a derogatory manner. Previ-

ously, abusive content classification in Tamil and Telugu languages are explored as Multi-class classification problems with the labels Homophobia, Misandry, Counter-speech, Misogyny, Xenophobia, and Transphobic in the shared tasks on RANLP-2023 and ACL-2022(Priyadharshini et al., 2023, 2022). The current dataset includes Tamil and Malayalam text, often containing code-mixed content. It is framed as a binary classification problem to detect the presence of abusive content targeting women with the labels "Abusive" and "Non-Abusive".

2.3 Fake News Detection (fake-news)

This task aims to identify fake news in Malayalam texts. Given the rapid spread of misinformation, the ability to detect fake news in regional languages is crucial for maintaining information integrity. The shared task consists of two datasets (Subramanian et al., 2024, 2025), one with binary classification labels as "Fake" and "Original"(**Task A**)(Subramanian et al., 2023) and the other dataset with multi-class classification labels as "Half-True", "False", "Partly-False" and "Mostly-False"(**Task B**)(Devika et al., 2024). Only the binary classification **Task A** is explored in this work.

3 Methodology

3.1 Data Preprocessing

The text data was preprocessed using the XLM-Roberta tokenizer to generate token embeddings and attention masks. The Multilingual MiniLM model uses the XLM-Roberta tokenizer as the former is a distilled version of the later model. Tokens were truncated and padded to a maximum length of 256. Labels were encoded as binary values for each task(Table 1).

Task	Negative (0)	Positive (1)
ai-gen-review	HUMAN	AI
abusive-woman	Non-Abusive	Abusive
fake-news	Original	Fake

Table 1: Task and Label Mapping

3.2 Model

The pre-trained checkpoint from Hugging Face, `microsoft/Multilingual-MiniLM-L12-H384`, is used as the base model for fine-tuning.

The MiniLM model architecture consists of 12 hidden layers, each with a hidden layer size of 384, totaling 21M[1] parameters.

3.3 Fine-Tuning

A classification head with a fully connected layer and softmax activation function was added on top of the base Multilingual MiniLM model using the `AutoModelForSequenceClassification` class from the transformers library by Hugging Face(Wolf et al., 2020). The model is trained using the `Trainer` API from the transformers library. Three models **ai-gen-review**, **abusive-woman** and **fake-news** were created as the result of the fine-tuning. Both the Tamil and Malayalam datasets are jointly used to fine-tune the models **ai-gen-review** and **abusive-woman** and the Malayalam dataset is used for fine-tuning the model **fake-news**. The best model was selected based on the f1-score evaluated during fine-tuning.

Models	Batch Size	No. of Epochs
ai-gen-review	128	6
abusive-woman	128	6
fake-news	256	9

Table 2: Hyperparameter configuration used for fine-tuning the model on different tasks.

4 Results

The fine-tuned models are then evaluated in Tamil and Malayalam for different tasks using F1-Score with macro averaging as the evaluation metric. The evaluation results on Tamil and Malayalam language Tasks are presented in Table 3 and Table 4 respectively.

The fine-tuned checkpoints of the models for AI-generated product review detection (Tamil & Malayalam), abusive text targeted at woman detection (Tamil & Malayalam) and Fake News Detection (Malayalam) are made available as a collection in Hugging Face[2].

Task	F1-Score	Rank	v/s Best
ai-gen-review	0.6745	24	-0.2955
abusive-woman	0.7821	5	**-0.0062**

Table 3: Evaluation Metrics for Tamil Tasks.

[1]This only includes the transformer parameters and does not include the embedding parameters

[2]https://huggingface.co/collections/livinNector/multilinugal-minilm-dravidianlangtech-679b3d894e207e2844c4d637

Task	F1-Score	Rank	v/s Best
ai-gen-review	0.8997	6	**-0.0202**
abusive-woman	0.6915	7	**-0.0656**
fake-news	0.8320	11	**-0.0660**

Table 4: Evaluation Metrics for Malayalam Tasks.

5 Investigating the results of AI-Generated Review detection in Tamil

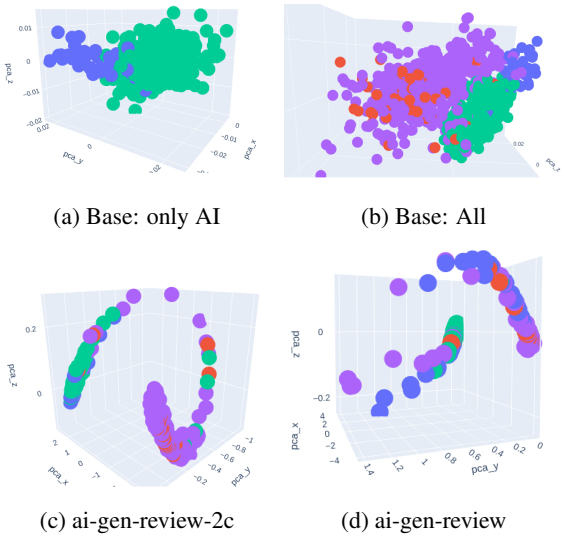
After the declaration of the result of the shared tasks, the reason for the significant variation in the performance of the **ai-gen review** model in the Tamil language test data is explored. Even though the model achieved an F1 score of 0.9816 in the development set during the initial fine-tuning, it had a lower F1 score in the test set. Three more fine-tuning runs are conducted with the same dataset, despite the F1 score with the development set (Tamil-Malayalam) being consistent in the range of 0.97-0.98, the F1 score in the test set varied significantly for the Tamil language, although the development set had a consistently high F1 score (Table 5). These different fine-tunings of the Multilingual MiniLM are available in Hugging Face.

Model	Tam-Test	Mal-Test
ai-gen-review	**0.6745**	0.8997
ai-gen-review-2a	0.8996	0.9147
ai-gen-review-2b	0.9095	0.8942
ai-gen-review-2c	**0.9800**	0.8749

Table 5: Macro F1-score evaluated on the Tamil and Malayalam test sets individually.

To study this in detail, a 3-dimensional visualization of the embedding space of the Multilingual MiniLM model is created using PCA transformation of the pooling layer outputs to study the embeddings of reviews in train and test sets. The visualizations show that the embeddings of AI-generated text in both train and test sets varied significantly. Also, visualizing the embeddings of the **ai-gen-review** and **ai-gen-review-2c** suggests that the model is capable of achieving a higher performance. The visualization of the embeddings is presented in Figure 1.

These explorations make it clear that the train and test set vary significantly in the embedding space of AI-generated reviews. This makes the training set insufficient to capture the entire region of AI-generated reviews. To overcome this issue in the data, more diverse AI-generated reviews from different AI models can be added to the training set so that the training and evaluation objectives of AI-generated reviews align closer and can be captured by the model.



(a) Base: only AI (b) Base: All

(c) ai-gen-review-2c (d) ai-gen-review

Figure 1: 3d visualization of PCA transformation of pooling layer outputs of the base and fine-tuned MiniLM models. green=AI-Train, blue=AI-Test, violet=HUMAN-Train, red=HUMAN-Test.

6 Conclusion

The results indicate that the performance of Multilingual MiniLM on the downstream tasks AI-generated review detection, abusive text detection, and fake news detection is comparable to the other models while having significantly fewer parameters than the other BERT-based models. The misalignment in the train and test sets of the Tamil AI-generated review data set is identified. The results of the fine-tuned models are also justified using visualization of the output layer.

7 Limitations

This study has the following limitations.

- The AI-generated review detection task exhibited significant variability in model performance, particularly for Tamil, suggesting limitations in the training dataset's representativeness.
- Though this study compares the performance of this model to the best performance reported by others from the shared task, it doesn't compare it with the performance of a model with a

similar architecture and more parameters like XLM-Roberta.

- Explainability and interpretability of the model's performance are not analyzed in detail. A detailed study on the intermediate attentions of the model might give hints on the tokens that contribute more to the results.
- The effects of different data augmentation and regularization techniques on the performance of the model have not been explored in this work.

References

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Asha Hegde, Kavya G, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2023. MUCS@DravidianLangTech2023: Leveraging learning models to identify abusive comments in code-mixed Dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 266–274, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Zhipeng Luo and Jiahui Wang. 2023. DeepBlueAI@DravidianLangTech-RANLP 2023. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 171–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_Punctuators@DravidianLangTech-EACL2024: Transformer-based approach for

detection and classification of fake news in Malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186, St. Julian's, Malta. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

CUET-823@DravidianLangTech 2025: Shared Task on Multimodal Misogyny Meme Detection in Tamil Language

Arpita Mallik, Ratnajit Dhar, Udoy Das[†], Momtazul Arefin Labib, Samia Rahman, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
[†]East Delta University, Bangladesh
{u2004023, u2004008}@student.cuet.ac.bd, udoy.d@eastdelta.edu.bd,
{u1904111, u1904022}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Misogynous content on social media, especially in memes, present challenges due to the complex reciprocation of text and images that carry offensive messages. This difficulty mostly arises from the lack of direct alignment between modalities and biases in large-scale visio-linguistic models. In this paper, we present our system for the Shared Task on Misogyny Meme Detection - DravidianLangTech@NAACL 2025. We have implemented various unimodal models, such as mBERT and IndicBERT for text data, and ViT, ResNet, and EfficientNet for image data. Moreover, we have tried combining these models and finally adopted a multimodal approach that combined mBERT for text and EfficientNet for image features, both fine-tuned to better interpret subtle language and detailed visuals. The fused features are processed through a dense neural network for classification. Our approach achieved an F1 score of 0.78120, securing 4th place and demonstrating the potential of transformer-based architectures and state-of-the-art CNNs for this task.

1 Introduction

The concept of memes have evolved into a powerful form of cultural transmission on the internet. Despite being widely debated in academic circles, memes have been adopted extensively, as evidenced by a surge of interest since 2011 and over 1.9 million Google search results for 'Internet meme,' which has highlighted their great cultural significance (Shifman, 2013).

Although memes have often been viewed as harmless entertainment, the internet has become a space where the harassment of women and marginalized groups has been documented widely in both academic and popular press. Feminist research has revealed that online sexism and harassment have frequently been reframed as "acceptable" form of humor (Drakett et al., 2018). The need to address multimodal issues has been highlighted by the fact that hateful content targeting women is not only found in text but also in visual, audio, or combined forms (Singhal et al., 2022).

The purposive focus of this paper has been to detect misogyny in Tamil-English code-mixed memes on social media platforms. The DravidianLangTech@NAACL 2025 conference (Chakravarthi et al., 2025) has introduced a dataset under the Shared Task on Misogyny Meme Detection (Ponnusamy et al., 2024), which has consisted of multimodal content with textual and visual elements labeled as misogynistic or non-misogynistic, enabling the identification of misogyny in a multimodal context.

To achieve our goal, we have applied various image and text augmentation techniques, such as brightness adjustment, grayscale transformations, and back-translation, and have evaluated various models, including transformer-based models (mBERT, IndicBERT), image models (ResNet, ViT, and EfficientNet), and multimodal models combining mBERT with EfficientNet, mBERT with ViT, and IndicBERT with ResNet. The results have shown that the multimodal models have outperformed the unimodal ones. With an F1 score of 0.78120, our approach has ranked 4th in the competition. The main contributions of this work have been:

- We have made comparison of various multimodal and unimodal models, to find the best suitable one for the given dataset.
- We have applied different types of data augmentation techniques, including back-translation, to address class imbalance.
- We have developed a preprocessing pipeline involving text transliteration and image enhancements to improve data quality.

Further implementation details can be accessed via the GitHub repository: [1].

2 Related Work

In the existing research on the topic of misogyny meme detection, only a few approaches have specifically addressed the issue, which is a major concern on social media platforms all over the world. Prior studies in this domain can be categorized based on their approach—text-based, image-based, and multimodal—as well as their use of machine learning, deep learning, and pre-trained models. Additionally, research efforts have focused on binary vs. multi-label classification and bias mitigation.

Nozza et al. (2021) has shown that hate speech detection models are not effective across different types of hate speech targets, highlighting the need for specialized approaches and datasets for detecting misogyny.

Recent studies have explored multimodal techniques to improve classification accuracy. For instance, Shaun et al. (2024) proposed a multimodal method for classifying Tamil and Malayalam memes as "Misogynistic" or "Non-Misogynistic", using Multinomial Naive Bayes, with outputs combined using weighted probabilities. This proved the effectiveness of combining modalities for low-resource languages.

Building on the success of prompt learning in NLP, some researchers have investigated prompt-based approaches for identifying harmful memes (Jindal et al., 2024). These methods involved converting images into textual representations to reduce the semantic gap between the text and images in memes.

Other studies have used advanced fusion techniques. Attanasio et al. (2022) presented a system using Perceiver IO for late fusion in misogynous meme detection. It combines ViT for images and RoBERTa for text, handling both binary and multi-label classification. The approach outperformed baseline models and showed the effectiveness of Perceiver IO for multimodal fusion. Hakimov et al. (2022) proposed a pre-trained CLIP model to extract text and image features, which are then combined with an LSTM layer. Pramanick et al., 2021 also used CLIP embeddings, along with intra-modality attention and cross-modality fusion in their proposed model MOMENTA, which is a multimodal deep learning model for detecting harmful memes along with their targets.

3 Data

We have utilized the dataset provided under the Shared Task on Misogyny Meme Detection - DravidianLangTech@NAACL 2025 (Ponnusamy et al., 2024). The dataset has been segmented into training, development, and test sets containing 1136, 284, and 356 samples, respectively. It primarily consists of code-mixed Tamil-English memes, the type of language commonly observed in online communication. The dataset consists of significantly lower number of misogynistic memes compared to non-misogynistic ones, as shown in Table 1.

Sets	Misogyny	Not-misogyny	Total
Train	285	851	1,136
Development	74	210	284
Test	89	267	356
Total	448	1,328	1,776

Table 1: Label Distribution for Misogyny Meme Detection Dataset.

4 Methodology

4.1 Data Preprocessing

In terms of image preprocessing, all images have been resized to a consistent dimension of $224 \times 224 \times 3$ pixels. Contrast and brightness have also been adjusted with specific control parameters to enhance image quality.

For text preprocessing, unwanted symbols, punctuation, numbers, URLs, and emojis have been removed. Tamil stopwords have been filtered out to preserve meaningful content. Since the dataset has comprised text in Tamil, English, and Tamil written in English script, the preprocessed text has been transliterated into Tamil using the Indic Transliteration library[2].

4.2 Data Augmentation

Targeted data augmentation techniques have been applied to tackle the class imbalance in our dataset. Since the misogynistic memes have been significantly fewer than the non-misogynistic ones, as

[1] https://github.com/ratnajit-dhar/CUET-823-Multimodal-Misogyny-Meme-Detection

[2] https://pypi.org/project/indic-transliteration/

reflected in Table 1, focus has been placed solely on augmenting the misogynistic memes to balance the dataset.

For image data, the torchvision library[3] has been used to apply brightness adjustment, grayscale transformation, and posterization. For text, back-translation has been employed using the deep-translator library[4] with two pipelines: Tamil → English → Tamil and Tamil → Malayalam → Tamil. Additionally, synonym replacement and paraphrasing augmentation techniques have been experimented with, but these methods have not yielded satisfactory results. Table 2 shows the class distribution after data augmentation technique has been applied.

Class	Before	After
Misogynistic	285	855
Non-Misogynistic	851	851

Table 2: Dataset distribution before and after augmentation.

4.3 Overview of Experimented Models

4.3.1 Unimodal model

We have fine-tuned mBERT and IndicBERT for textual features, tokenizing with a max length of 512 tokens. The models have been trained on 20 topics with a batch size of 8, using the AdamW optimizer and a learning rate of 2e-5.

For image data, we have experimented with Vision Transformers (ViT), ResNet, and EfficientNet. The images were converted to PIL format, resized to 224×224, converted to tensors, and normalized using standard ImageNet values. The models have been trained for 20 epochs with a batch size of 16, using the Adam optimizer with a learning rate of 1e-4. The final fully connected layer has been replaced to match the number of unique classes. The architecture of the unimodal text model is shown in Figure 1, while the unimodal image model is illustrated in Figure 2

4.3.2 Multimodal model

Building on these unimodal baselines, we have explored multimodal models, such as mBERT with ResNet, mBERT with ViT, and IndicBERT with EfficientNet. These multimodal architectures allow the model to learn complex interactions between

[3] https://pytorch.org/vision/stable/index.html
[4] https://pypi.org/project/deep-translator/

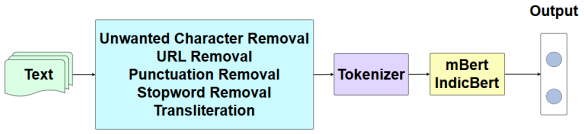


Figure 1: Unimodal Architecture for Text Data Processing and Classification.

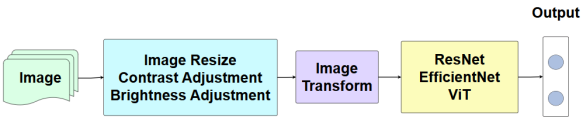


Figure 2: Unimodal Architecture for Image Data Processing and Classification.

textual and visual inputs, as shown in Figure 3. Our best-performing multimodal model, mBERT + EfficientNet, was fine-tuned using a 512-sequence-length mBERT tokenizer and images resized to 224×224 pixels. The 768-dimensional text embeddings and 1280-dimensional image features have been combined via a fully connected layer. The model was trained for 20 epochs with a batch size of 16, and the learning rate was set to 1e-4. For the mBERT + ViT model, text and image features were concatenated using 768 dimensions from both modalities. In the IndicBERT + ResNet model, the concatenation involved 768-dimensional text embeddings and 2048-dimensional image features.

5 Results and Analysis

This section presents the outcomes of our misogyny meme classification task, comparing unimodal and multimodal approaches to highlight their effectiveness in addressing classification challenges.

We have evaluated the performance of our models using weighted precision, recall, F1 score, and macro-averaged F1 score (Macro-F1). The Macro-F1 score is considered as the primary metric for assessing the final performance of the systems.

5.1 Comparative Analysis

We have assessed the performance of various models and found that among the unimodal text classifiers, mBERT performed better than IndicBERT with a higher F1 score of 0.69 compared to 0.62. For unimodal image classifiers, ResNet achieved a better score than EfficientNet and ViT.

When we combined mBERT with EfficientNet in a multimodal setup, the model achieved the highest F1 score of 0.78 with a precision of 0.80 and a recall of 0.77. This shows the strength of com-

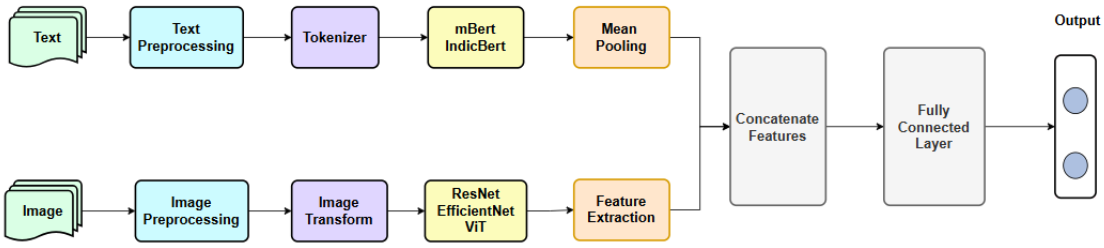


Figure 3: Multimodal Architecture for Text-Image Classification.

bining textual and visual modalities for improved performance.

We have also evaluated the performances of mBERT + ResNet, IndicBERT + ViT, and various other combinations of multimodal models. However, Table 3 shows only the three best performing approaches.

5.2 Error Analysis

To further analyze model performance, we present the confusion matrix in Figure 4. It shows that our best multimodal model (mBERT + EfficientNet) correctly classified 247 non-misogynistic and 54 misogynistic memes. However, it misclassified 20 non-misogynistic memes as misogynistic (false positives) and 35 misogynistic memes as non-misogynistic (false negatives). This suggests that the model may sometimes rely on certain words or patterns rather than the overall meaning, leading to incorrect predictions.

	Classifier	Macro Average		
		P	R	F1
Unimodal (Text)	mBert	0.68	0.71	0.69
	IndicBert	0.64	0.61	0.62
Unimodal (Image)	ResNet	0.77	0.74	0.74
	EfficientNet	0.76	0.69	0.72
	ViT	0.71	0.63	0.65
Multi-modal	mBert+ EfficientNet	**0.80**	**0.77**	**0.78**
	mBert+ViT	0.79	0.72	0.75
	IndicBert+ ResNet	0.79	0.73	0.75

Table 3: Performance of different systems on the test dataset.

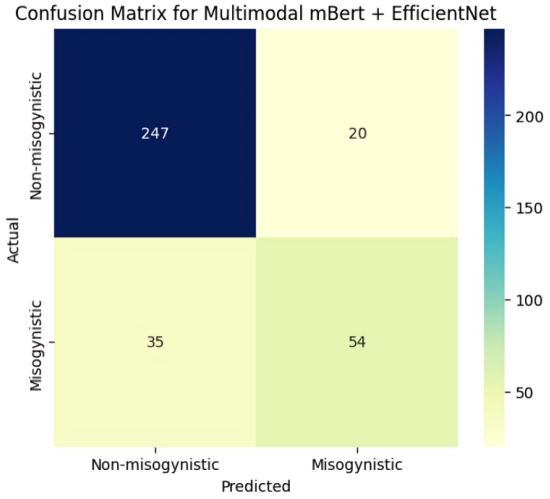


Figure 4: Confusion Matrix of the Multimodal mBert-EfficientNet Model.

6 Conclusion

In this research, we have evaluated various unimodal models and then compared various combinations of multimodal models for detecting misogynistic content in memes. Regardless of the challenges from the limited dataset, multimodal models have surpassed unimodal approaches, highlighting the importance of incorporating textual and visual information. We have discovered that the multimodal mBert + EfficientNet has performed the best among the other multimodal approaches, with an F1 score of 0.78. Future work will focus on expanding the dataset, improving data augmentation techniques, and better utilizing both text and images to detect subtle misogynistic content more effectively.

Limitations

The dataset provided for our task has been relatively small and imbalanced, with an inadequate

number of misogynistic memes. Although data augmentation techniques have been applied, the dataset size has still impacted performance, especially for the minority class. Additionally, although contrast and brightness adjustments were applied to improve image quality, but they could not fully eliminate noise and inconsistencies, leading to some misclassifications. Lastly, our model's performance could be further improved by training on more nuanced and subtle examples, which are currently underrepresented in the training data.

Ethics Statement

In this study, we have developed our methodology following the highest ethical practices. By contributing to the identification of misogynistic content in Tamil-English code-mixed memes, we hope to make the internet a safer and more inclusive place. We are committed to sharing our findings to prevent online misogyny while respecting linguistic and cultural diversity.

References

Giuseppe Attanasio, Debora Nozza, Federico Bianchi, et al. 2022. Milanlp at semeval-2022 task 5: Using perceiver io for detecting misogynous memes with text and image modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Jessica Drakett, Bridgette Rickett, Katy Day, and Kate Milnes. 2018. Old jokes, new media–online sexism and constructions of gender in internet memes. *Feminism & psychology*, 28(1):109–127.

Sherzod Hakimov, Gullal S Cheema, and Ralph Ewerth. 2022. Tib-va at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. *arXiv preprint arXiv:2204.06299*.

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Debora Nozza, Federico Bianchi, Dirk Hovy, et al. 2021. Honest: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

H Shaun, Samyuktaa Sivakumar, R Rohan, Nikilesh Jayaguptha, and Durairaj Thenmozhi. 2024. Quartet@ lt-edi 2024: A svm-resnet50 approach for multitask meme classification-unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 221–226.

Limor Shifman. 2013. *Memes in digital culture*. MIT press.

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 1322–1331.

Hermes@DravidianLangTech 2025: Sentiment Analysis of Dravidian Languages using XLM-RoBERTa

Emmanuel George P[1], **Ashiq Firoz**[1], **Madhav Murali**[1]
Siranjeevi Rajamanickam[2] , **Balasubramanian Palani**[3]
[1]Department of Computer Science and Engineering, IIIT Kottayam
[2]Lecturer, Dept of Computer Engineering, Govt. Polytechnic College-Trichy
[3]Assistant Professor, Indian Institute of Information Technology Kottayam
{emmanuel22bcs104,ashiq22bcd13,madhav22bcs50,pbala}@iiitkottayam.ac.in,rajasiranjeevi@gmail.com

Abstract

Sentiment analysis, the task of identifying subjective opinions or emotional responses, has become increasingly significant with the rise of social media. However, analyzing sentiment in Dravidian languages such as Tamil-English and Tulu-English, presents unique challenges due to linguistic code-switching (where people tend to mix multiple languages) and non-native scripts. Traditional monolingual sentiment analysis models struggle to address these complexities effectively. This research explores a fine-tuned transformer model based on the XLM-RoBERTa model for sentiment detection. It utilizes the tokenizer from the XLM-RoBERTa model for text preprocessing. This research is based on our work for the Sentiment Analysis in Tamil and Tulu DravidianLangTech@NAACL 2025 competition. We received an F1-score of 71% for the Tulu dataset and 60% for the Tamil dataset, which placed us third in the competition.

1 Introduction

Sentiment analysis plays a pivotal role in understanding subjective opinions and emotional responses in text. With the growing prominence of social media, the demand for sentiment analysis on user-generated content has surged. However, social media texts often include code-mixed content, where multiple languages are blended within a single sentence. This phenomenon is particularly prevalent in multilingual communities, such as those speaking Dravidian languages, where Tamil-English and Tulu-English code-mixing is common. These texts often incorporate linguistic code-switching and non-native scripts, adding layers of complexity to the task of sentiment analysis.

The novelty of this research lies in its focus on sentiment analysis for low-resource code-mixed Dravidian languages, an area that has remained underexplored despite the growing demand for multilingual natural language processing (NLP) solutions. Additionally, this study is based on a fine-tuned XLM-RoBERTa (Liu et al., 2019) model, providing new insights into their strengths and limitations in handling linguistic complexities like code-switching and non-native scripts. The fine-tuned transformer approach, in particular, demonstrates a superior ability to address these challenges by effectively leveraging its contextual understanding capabilities.

The dataset (Chakravarthi et al., 2020b)used in this study contained four and five sentiment classes in Tamil and Tulu language, respectively. The XLM-RoBERTa model (Conneau et al., 2020)was trained using these datasets for the multi-class classification task. Later the macro F1-score metric of the validation dataset was used to evaluate the model performance. This approach secured a third-place ranking in the competition(Durairaj et al., 2025).

A comparison of the results of XLM-RoBERTa model with tradition machine learning models like Logistic Regression and Random Forest in combination with the TF-IDF vectorizer for tokenization was done and this comparison was extended to the transformer models BERT base (Devlin et al., 2019) and RoBERTa base (Palani and Elango, 2023). The XLM-RoBERTa model provided a better result in comparison with all these models also.

2 Literature Survey

Ahmad et al. (2022) (Ahmad et al., 2022) provide a comprehensive review of machine learning techniques for sentiment analysis in code-mixed and switched text, emphasizing the challenges posed by bilingual and multilingual expressions in Indian social media contexts. Chakravarthi et al. (2020) (Chakravarthi et al., 2020a) introduced a gold standard corpus for Malayalam-English code-mixed text, which serves as a benchmark for sentiment

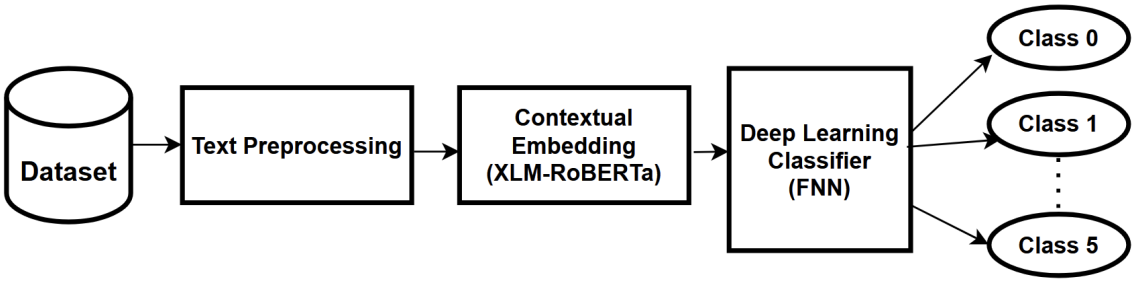


Figure 1: Architecture of the proposed model for sentiment prediction

analysis tasks with high inter-annotator agreement. The Language Technologies Research Center at IIIT Hyderabad studied (Mishra et al., 2018) sentiment detection in Hindi and Bengali using a voting classifier with an SVM model and TF-IDF vectorization, achieving accuracies of 56% for Hindi and 52% for Bengali. RANLP 2023 explored sentiment analysis for code-mixed Tamil and Tulu texts, (Hegde et al., 2023) reporting macro-average F1 scores of 0.32 for Tamil and 0.542 for Tulu, underscoring the growing interest in Dravidian sentiment analysis and the need for further advancements.

3 Methodology

We followed a series of steps to develop the model for sentiment prediction, which included dataset preprocessing, tokenization, training, validation, prediction, and model evaluation. Each of these steps is discussed in detail in this section, with visual representation provided in Figure 1.

3.1 Problem Definition

The sentiment analysis task involves two distinct datasets: one for Tamil and another for Tulu. Given a Tamil dataset $T = \{t_1, t_2, \ldots, t_m\}$ consisting of m social media comments, each comment $t_i \in T$ is associated with one of the following class labels: Positive, Negative, Mixed Feelings, or Unknown State. Separately, the Tulu dataset $U = \{u_1, u_2, \ldots, u_k\}$ consists of k social media comments, each labeled with one of five categories: Not Tulu, Positive, Negative, Neutral, or Mixed. Classification models $f_T : T \rightarrow y$ and $f_U : U \rightarrow y$ are defined and trained to predict the corresponding class label for each comment in their respective datasets.

3.2 Data Preprocessing

Data processing prepares the dataset for training the machine learning model to detect the different sentiments.The dataset had multiple entries with Nan values in it. We had to remove these values from the dataset and we went on to map the class labels to the integer values using a label map.

3.3 Tokenization

Converted the textual data to numbrs using the process of tokenization and the tokenizer we used was the XLM-RoBERTa tokenizer which would align perfectly with our classifier model. Similarly the RoBERTa base and BERT base models (Palani and Elango, 2023) utilized their respective tokenizers and the machine learning models utilized the TF-IDF vectorizer (Kanta and Sidorov, 2023) with max_features set to 5000.

3.4 Model Architecture

The model we used here is the XLM-RoBERTa (Conneau et al., 2020) model and feed forward networks(FFN). We set the problem_type parameter to 'single_label_classification,' meaning that each data point will be assigned to only one of the target classes and num_labels to the number of classes in the dataset. (Liu et al., 2019) Both RoBERTa base and BERT base(Devlin et al., 2019) where also set to the same parameters. Incase of Logistic Regression it had max_iter set to 1000 and for Random Forest n_estimators was set to 200 .

3.5 Model Training

The model was trained using the AdamW optimizer (learning rate: 2 x 10^{-5}, weight decay: 0.01) with a linear warm-up schedule. A batch size of 16 was used, and training ran for 10 epochs with early stopping (patience: 3). Cross-Entropy Loss was

Table 1: Samples of different class labels in the dataset

Tamil		Tulu	
Positive	2020 முதல் வெற்றி மாஸ் வெறித்தனமான	Positive	ಅಣ್ಣ ಮಸ್ತ್ ಖುಷಿ ಆಪುಂಡು ಇರೆನ ಶೋ ತೂವರೆ
Negative	திமிர் பிடித்த திருநங்கைகள்	Negative	ಎಂಚಿ ಸಾವುದ ಪುಕುಳಿಯ
Mixed_feelings	சில்லறை தராமல் எடுத்துக் கொண்டு போனார்	Mixed	ಅಂಬಾನಿ ತುಂಡಾ ಸೈತೆ ಪೊವೇ
unknown_state	இந்த படத்தை டிவில பத்து டே 2022 ல	Neutral	ನೀರ್ ದ ಮಹತ್ವನ್ ತೇರಿಲೆ
-	-	Not Tulu	Congratulations Mohan sir

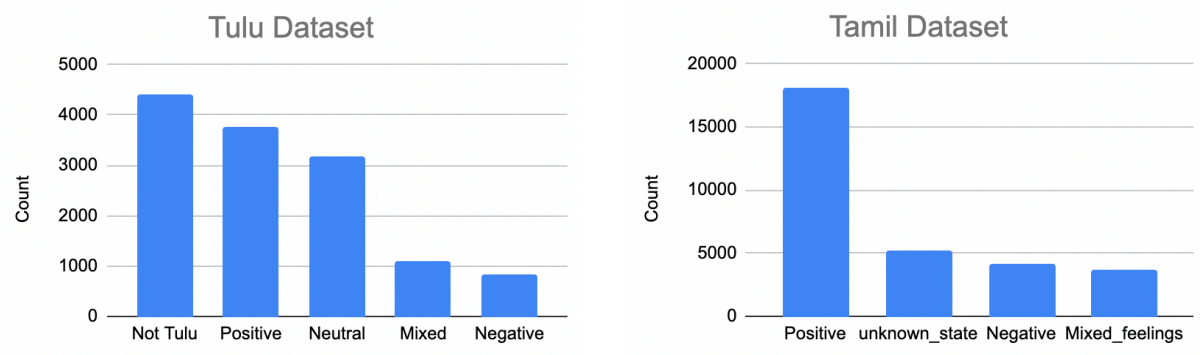


Figure 2: Frequency distribution of classes in the training dataset

applied, and overfitting was monitored using the weighted average F1-score on the test set. Transformer models followed a similar training setup, while machine learning models used the fit function.

3.6 Model Evaluation

The model was evaluated using the validation dataset. For each datapoint, the model predicted that it belonged to a predefined class label. Now, the different evaluation metrics were found and compared on the basis of the models.

4 Experiment

This section gives us a comprehensive study of the experimental setup and the different performance metrics utilized in this study.

4.1 Experimental Setup

The study uses machine learning and deep learning approaches to classify sentiments. These models were implemented and tested in Python programming language using PyTorch, Transformers, Pandas, Scikit-learn, and Tqdm for data processing, model training, and analysis. Hugging Face tokenizers handled text preprocessing and training was performed on CUDA-enabled GPUs, leveraging PyTorch for efficient deep learning and Hugging Face API for streamlined access to the models.

4.2 Dataset Description

This study utilizes multilingual sentiment analysis datasets in Tamil and Tulu, including cleaned subsets of Tamil data. The dataset(Hegde et al., 2022) contains training and validation sets and the frequency distribution of different class labels in the training dataset are shown in Figure 2. Some samples of each of these class labels are also shown in Table 1 and a summary of the datasets is shown in Table 2.

The dataset incorporates both code-mixed and romanized data points and samples of code-mixed and romanized data points are as given below:

- Code-mixed: படம் வெற்றிபெற நாடார் சமுகத்தினர் சார்பாக வாழ்த்துகள்
- Romanized: kandipa nama Ella records um break panuvom

Table 3: Performance comparison of XLM-RoBERTa model with other standard models on Tamil and Tulu datasets

Model	Tulu Dataset				Tamil Dataset			
	Acc.	Prec.	Rec.	Macro F1	Acc.	Prec.	Rec.	Macro F1
XLM-RoBERTa	0.71	0.63	0.59	0.59	0.65	0.51	0.49	0.49
RoBERTa base	0.67	0.58	0.55	0.54	0.63	0.48	0.43	0.44
BERT base	0.69	0.59	0.59	0.59	0.64	0.49	0.48	0.48
TF-IDF (LR)	0.64	0.55	0.57	0.55	0.53	0.44	0.48	0.45
TF-IDF (SVM)	0.64	0.54	0.51	0.51	0.63	0.49	0.39	0.41

Table 2: Summary of datasets

Dataset	Training	Testing
Tamil	31,122	3,843
Tulu	13,308	1,643

4.3 Evaluation Metrics

The performance of the model is evaluated using Accuracy, Macro Precision, Macro Recall, and Macro F1-score.

$$\text{Accuracy (Acc.)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Macro Precision (Prec)} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i} \quad (2)$$

$$\text{Macro Recall (Rec.)} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \quad (3)$$

$$\text{Macro F1} = \frac{2 \times \text{Macro Precision} \times \text{Macro Recall}}{\text{Macro Precision} + \text{Macro Recall}} \quad (4)$$

Here, TP, TN, FP, and FN represent the true positives, true negatives, false positives, and false negatives, respectively, and N is the number of classes.

5 Results

The best results for Tamil and Tulu sentiment analysis were given by the XLM-RoBERTa model, with a macro F1-score of 59% for the Tulu dataset and 49% for the Tamil dataset. XLM-RoBERTa model was actually trained on multi-lingual text and this is the reason why it has an edge over the other models. The details of the macro of the evaluation metrics and accuracy are as shown in the Table 3. The best macro average of f1-score we received was 59% and 49% for testing dataset of Tulu and Tamil language respectively.

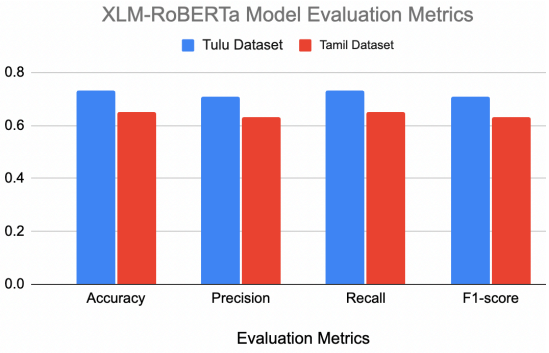


Figure 3: Evaluation metrics of XLM-RoBeRTa model.

From Figure 3 we can infer that the Tulu dataset is performing better than the Tamil dataset, this is because of the irregular frequency of the class labels in the Tamil dataset. As you can see in the Figure 2 the number of positive classes in the dataset is very high compared to the rest of the classes causing class imbalance thus leading to lesser accuracy.

6 Conclusion and Future Directions

Sentiment analysis in code-mixed text faces challenges like class imbalance and code-switching complexities, with fine-tuned transformers helping but limited by multilingual resource gaps. The Tamil dataset shows significant class imbalance, affecting model performance, with XLM-RoBERTa achieving a macro F1-score of 59% for Tulu and 49% for Tamil. Future work could explore re-sampling, cost-sensitive learning, and advanced data augmentation to address imbalance, while developing comprehensive lexical resources would enhance transformer-based sentiment analysis for multilingual communities.

References

Gazi Ahmad, Jimmy Singla, Anis Ali, Aijaz Reshi, and Anas A. Salameh. 2022. Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus - a comprehensive review. *International Journal of Advanced Computer Science and Applications*, 13.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, Lavanya S K, Thenmozhi D., Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in Tamil and Tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Selam Kanta and Grigori Sidorov. 2023. Selam@DravidianLangTech:sentiment analysis of code-mixed Dravidian texts using SVM classification. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 176–179, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Pruthwik Mishra, Prathyusha Danda, and Pranav Dhakras. 2018. Code-mixed sentiment analysis using machine learning and neural network approaches. *Preprint*, arXiv:1808.03299.

Balasubramanian Palani and Sivasankar Elango. 2023. Ctrl-fnd: content-based transfer learning approach for fake news detection on social media. *International Journal of System Assurance Engineering and Management*, 14(3):903–918.

SSNTrio@DravidianLangTech 2025: Identification of AI Generated Content in Dravidian Languages using Transformers

Bhuvana J
Sri Sivasubramaniya Nadar College of Engineering
bhuvanaj@ssn.edu.in

Mirnalinee T T
Sri Sivasubramaniya Nadar College of Engineering
MirnalineeTT@ssn.edu.in

Rohan R
Sri Sivasubramaniya Nadar College of Engineering
rohan2210124@ssn.edu.in

Diya Seshan
Sri Sivasubramaniya Nadar College of Engineering
diya2210208@ssn.edu.in

Avaneesh Koushik
Sri Sivasubramaniya Nadar College of Engineering
avaneesh2210179@ssn.edu.in

Abstract

The increasing prevalence of AI-generated content has raised concerns about the authenticity and reliability of online reviews, particularly in resource-limited languages like Tamil and Malayalam. This paper presents an approach to the Shared Task on Detecting AI-generated Product Reviews in Dravidian Languages at NAACL2025, which focuses on distinguishing AI-generated reviews from human-written ones in Tamil and Malayalam. Several transformer-based models, including IndicBERT, RoBERTa, mBERT, and XLM-R, were evaluated, with language-specific BERT models for Tamil and Malayalam demonstrating the best performance. The chosen methodologies were evaluated using Macro Average F1 score. In the rank list released by the organizers, team SSNTrio, achieved ranks of 3rd and 29th for the Malayalam and Tamil datasets with Macro Average F1 Scores of 0.914 and 0.598 respectively.

1 Introduction

The swift progress in AI-generated text has sparked apprehensions regarding authenticity and dependability across multiple sectors, particularly in online reviews, where user-generated content plays a crucial role in shaping consumer choices. As AI technologies evolve, it has become increasingly difficult to differentiate between reviews authored by humans and those produced by AI. To maintain transparency on online platforms, it is essential to implement effective methods for detecting AI-generated content, especially in low-resource languages such as Tamil and Malayalam, which are characterized by a scarcity of annotated datasets and linguistic tools.

Identifying AI-generated content in Tamil and Malayalam poses distinct challenges due to their complex morphology, agglutinative structure, and varied writing styles. Standard multilingual models frequently struggle to grasp the linguistic nuances of these languages, highlighting the need for tailored approaches. Transformer based models, particularly those fine-tuned for Tamil and Malayalam, have demonstrated potential in overcoming these obstacles by developing context-aware representations suited to these languages.

This paper outlines a methodology for the Shared Task (Premjith et al., 2025) on Detecting AI-generated Product Reviews in Dravidian Languages, which centers on the classification of AI-generated versus human-written reviews in Tamil and Malayalam. The objective of the task was to assess the efficiency of various models in recognizing AI-generated text and to investigate the challenges that are specific to Dravidian languages.

2 Related Works

The detection of AI-generated text has become a critical area of research due to the rise of generative AI tools such as ChatGPT. Several studies (Mindner et al., 2023) have aimed to distinguish between human-written and AI-generated content, employing a range of features such as perplexity, semantic analysis, readability, and AI feedback. These features have been used to improve detection accuracy across various types of AI-generated texts. BERT-based models (Javaji et al., 2024), for example, have demonstrated impressive performance, achieving F1-scores over 96% for identifying basic AI-generated texts and over 78% for AI-rephrased texts. These models are particularly effective due to their ability to capture contextual and semantic nuances within text.

Other studies (Mozes et al., 2021) have explored the issue of adversarial attacks on natural language processing (NLP) models, revealing that humans are capable of generating adversarial examples through semantic-preserving word substitutions. These attacks have raised important questions about the vulnerability of NLP systems and the importance of validating adversarial inputs for

maintaining text integrity. Moreover, the identification of AI-generated content extends to broader concerns such as plagiarism, fake news, and content authenticity. For instance, large datasets consisting of essays have been employed to train machine learning models that classify content as either AI-generated or human-written. This approach (Wang et al., 2024) is particularly relevant in academic and online environments where ensuring the authenticity of text is paramount.

Additionally, efforts (Abburi et al., 2023) have been made to attribute specific language models to AI-generated text, as demonstrated by ensemble neural models that combine pre-trained LLMs to classify text. These advances reflect the growing importance of detecting AI-generated material in various domains, from education to online content moderation, and emphasize the need for robust and efficient detection systems.

It is clear that, Language-specific BERT models have not been utilized for this task. Leveraging the potential of such models by using their advanced architectures could serve as a novel contribution to this study.

3 Dataset Description

This task is focused on the precise identification of AI-generated reviews in Dravidian languages. It consists of two subtasks, differentiated by language: Tamil and Malayalam. The training dataset, as provided by the organisers, comprises the Review ID, the corresponding review text, and the associated label indicating whether the review is generated by AI or Human. Refer to Table 1 for detailed dataset statistics of each language.

	Tamil	Malayalam
Training Data	808	800
Testing Data	100	200

Table 1: Dataset Statistics

Figure 1 and Figure 2 show the distribution of labels in the Malayalam and Tamil datasets, respectively.

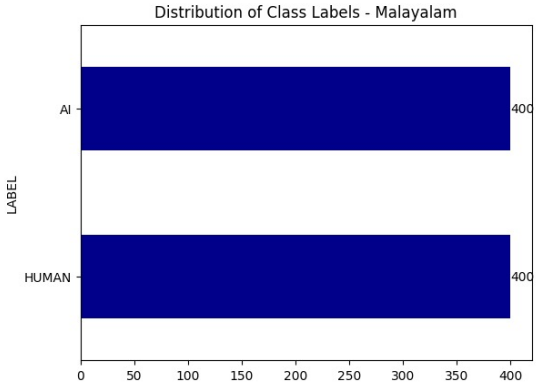


Figure 1: Label Distribution in Malayalam Dataset

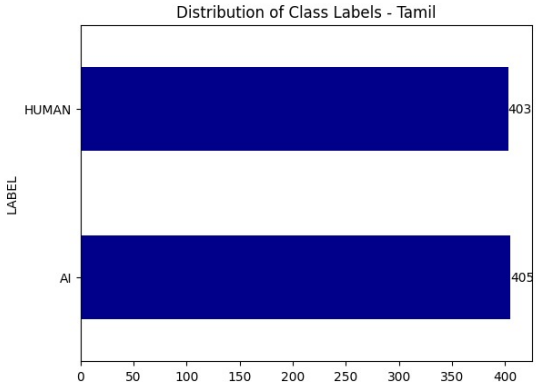


Figure 2: Label Distribution in Tamil Dataset

4 Methodology

4.1 Data Preprocessing

The methodology starts with preprocessing the dataset to ensure consistency and prepare the text for model training. Text data is converted to lowercase to maintain uniformity. Punctuation is removed to reduce noise, and the text is tokenized into individual words. Stopwords are removed to focus on meaningful terms, enhancing the quality of the textual data. The cleaned text is then reassembled for downstream processing. Additionally, the labels "AI" and "Human" are encoded to facilitate the model's understanding and classification tasks.

4.2 Synthetic Sample Generation

Recognizing the challenge of a limited dataset, SMOTE (Synthetic Minority Over-sampling Technique) upsampling was used to generate synthetic samples, enriching the dataset. This technique creates new synthetic examples by interpolating between existing samples, helping to enhance data diversity while preserving meaningful patterns. SMOTE mitigates class imbalance by ensuring that the classifier is exposed to a more balanced distribution of examples, reducing bias towards the majority class. Through this method, the dataset size was increased to 1,200 samples per label, improving the model's ability to generalize across all label categories and enhancing classification per-

formance.

4.3 Tokenisation and Feature Representation

For feature representation, the preprocessed text undergoes tokenization with the BERT tokenizer. The tokenizer transforms text input into numerical sequences, consisting of input IDs and attention masks, ensuring it works with the pre-trained BERT model. For consistency, padding and truncation are utilized, restricting the tokenized sequences to a maximum length of 128 tokens. This approach guarantees that the model can efficiently manage textual inputs of different lengths.

4.4 Model Building

After preprocessing and tokenization, the dataset was split into training and testing sets. The tokenizer was fit on the training data and applied to the test data to maintain consistency. Various pre-trained transformer models, including BART, IndicBERT, and RoBERTa, were tested, but Tamil BERT and Malayalam BERT performed best. These models likely excelled due to their language specialization, improving accuracy and contextual understanding.

The fine-tuned Tamil and Malayalam BERT models (Joshi, 2022) were trained on the respective datasets for 25 epochs, using the Adam optimizer (Zhang, 2018) and sparse categorical cross-entropy as the loss function. Both models were designed to handle binary classification tasks efficiently, leveraging their contextual embeddings to predict target labels accurately. This approach ensured robust performance while maintaining the interpretability and scalability of the solution.

5 Performance Metrics

This section provides insight on the metrics used to evaluate the performance of the methodologies employed for each task.

- **Recall** for a specific label indicates the proportion of correctly identified instances of that label out of all true instances in the data.
- **Precision** for a specific label measures the proportion of correctly identified instances of that type out of all instances predicted as that type by the model.
- **F1 Score** for a specific label is the harmonic mean of precision and recall for that label, providing a balanced measure.
- **Macro-Average F1 Score** evaluates the model's performance across all classes in multi-label classification by calculating the F1 score for each class independently, ensuring that each class is given equal weight in the overall score.

Language	F1 Score	Precision	Recall	Accuracy
Malayalam	0.91	0.92	0.92	0.92
Tamil	0.59	0.73	0.65	0.66

Table 2: Performance Scores

6 Result Analysis

Several transformer models were evaluated, including IndicBERT, RoBERTa, mBERT, XLM-R (Conneau et al., 2019), and others, for the task of distinguishing AI-generated content from human-written text in Tamil and Malayalam. The Malayalam and Tamil BERT models consistently demonstrated superior performance, proving to be the most effective for text classification in these languages.

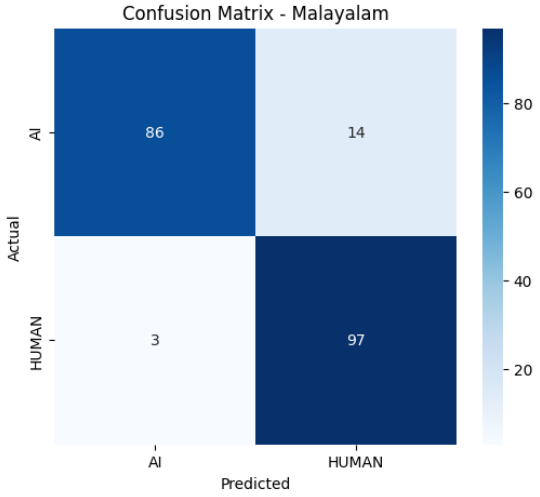


Figure 3: Confusion Matrix - Malayalam

Additionally, it was noticed that SMOTE upsampling played a crucial role in improving model performance by increasing the dataset size. This enhancement significantly boosted the model's ability to generalize across different categories, leading to a significant improvement in classification accuracy and overall robustness.

Using this model, test datasets were classified and key evaluation metrics were computed, including accuracy, precision, recall, and F1 scores for each language. The final performance results are presented in Table 2. The submissions ranked 3rd

for Malayalam and 29th for Tamil in the official ranklist released by the organizers.

The confusion matrices shown in Figures 3 and 4 depict the model's classification accuracy and the distribution of predicted labels compared to the actual labels for the Malayalam and Tamil test sets, respectively.

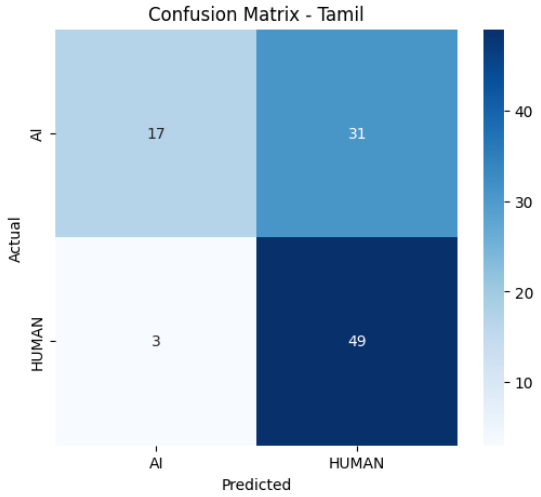


Figure 4: Confusion Matrix - Tamil

The model demonstrated high recall for Human text, indicating strong recognition of human-written content. However, there were challenges in correctly identifying AI text. This suggests that there may be feature overlap between AI and Human texts, and further model refinement is needed to improve AI content detection.

7 Conclusion

In conclusion, this paper has investigated the identification of AI-generated reviews for Indian languages, focusing mainly on Tamil and Malayalam.

The Malayalam and Tamil BERT models demonstrated superior performance for several reasons. First, these language-specific BERT models were fine-tuned on large corpora of Tamil and Malayalam texts, enabling them to better capture the unique syntactic and semantic nuances of these languages.

Second, the models' architecture was better suited to handle the linguistic characteristics of Tamil and Malayalam, which are significantly different from other languages, resulting in improved classification accuracy. This fine-tuning contributed to more precise identification of AI-generated content compared to other models.

The evaluation and analysis of the results from the tasks provided valuable insights into the challenges of Identifying and classification in multilingual contexts. Ongoing advancements in model refinement and feature extraction will be crucial for enhancing performance in future research efforts.

8 Future Enhancements

Neural networks can enhance AI-generated text detection by leveraging deep learning architectures. Transformer-based models like BERT, RoBERTa, and XLM-R can be fine-tuned for Tamil and Malayalam to improve classification accuracy. Hybrid models combining CNNs for feature extraction and LSTMs for sequential learning can further enhance performance. Additionally, explainable AI techniques like LIME and SHAP can provide insights into model decisions. Future work can also explore adversarial training and semi-supervised learning to improve robustness and generalization. Incorporating contrastive learning can help the model better distinguish between subtle differences in AI and human-generated text. Moreover, multi-modal approaches that integrate speech and text analysis could further improve detection accuracy in social media and real-world applications.

9 Limitations

Despite strong performance, the Tamil and Malayalam BERT models face limitations. Their effectiveness depends on the quality and balance of training data, with biases affecting classification accuracy. Additionally, the dataset may not fully represent real-world reviews, limiting generalization across domains.

A key challenge is the presence of domain-specific biases. Models trained on limited sources may struggle with sectors like healthcare or finance. Code-mixing and transliteration in Tamil and Malayalam further complicate classification, leading to errors.

Furthermore, the model's applicability to other Indian languages remains uncertain without fine-tuning. An important area for further exploration is error analysis, which could help identify issues related to ambiguous structures, data limitations, or model biases. Finally, as AI-generated content evolves to mimic human writing more closely, maintaining high classification accuracy will require periodic retraining with updated datasets and the incorporation of more advanced linguistic features.

References

Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. Generative ai text classification using ensemble llm approaches. *Preprint*, arXiv:2309.07755.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Prashanth Javaji, Pulaparthi Satya Sreeya, and Sudha Rajesh. 2024. Detection of ai generated text with bert model. In *2024 2nd World Conference on Communication Computing (WCONF)*, pages 1–6.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of human- and ai-generated texts: Investigating features for chatgpt. In *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, pages 152–170, Singapore. Springer Nature Singapore.

Maximilian Mozes, Max Bartolo, Pontus Stenetorp, Bennett Kleinberg, and Lewis D. Griffin. 2021. Contrasting human- and machine-generated word-level adversarial examples for text classification. *CoRR*, abs/2109.04385.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the shared task on detecting ai generated product reviews in dravidian languages: Dravidianlangtech@naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Hao Wang, Jianwei Li, and Zhengyu Li. 2024. Ai-generated text detection and classification based on bert deep learning algorithm. *Preprint*, arXiv:2405.16422.

Zijun Zhang. 2018. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee.

SSNTrio@DravidianLangTech 2025: Sentiment Analysis in Dravidian Languages using Multilingual BERT

Bhuvana J
Sri Sivasubramaniya Nadar College of Engineering
bhuvanaj@ssn.edu.in

Mirnalinee T T
Sri Sivasubramaniya Nadar College of Engineering
MirnalineeTT@ssn.edu.in

Diya Seshan
Sri Sivasubramaniya Nadar College of Engineering
diya2210208@ssn.edu.in

Rohan R
Sri Sivasubramaniya Nadar College of Engineering
rohan2210124@ssn.edu.in

Avaneesh Koushik
Sri Sivasubramaniya Nadar College of Engineering
avaneesh2210179@ssn.edu.in

Abstract

This paper presents an approach to sentiment analysis for code-mixed Tamil-English and Tulu-English datasets as part of the DravidianLangTech@NAACL 2025 shared task. Sentiment analysis, the process of determining the emotional tone or subjective opinion in text, has become a critical tool in analyzing public sentiment on social media platforms. The approach discussed here uses multilingual BERT (mBERT) fine-tuned on the provided datasets to classify sentiment polarity into various predefined categories: for Tulu, the categories were positive, negative, not_tulu, mixed, and neutral; for Tamil, the categories were positive, negative, unknown, mixed_feelings, and neutral. The mBERT model demonstrates its effectiveness in handling sentiment analysis for code-mixed and resource-constrained languages by achieving an F1-score of 0.44 for Tamil, securing the 6th position in the ranklist; and 0.56 for Tulu, ranking 5th in the respective task.

1 Introduction

Sentiment analysis involves identifying subjective opinions or emotional responses related to a specific topic. Over the past two decades, it has gained significant attention in both academia and industry. The demand for sentiment detection in social media texts, especially those containing code-mixing in Dravidian languages, has been steadily increasing.

The DravidianLangTech@NAACL 2025 shared task (Durairaj et al., 2025) focuses on analyzing sentiment in code-mixed Tamil-English and Tulu-English text collected from social media platforms like YouTube (S. K. et al., 2024). Both languages represent diverse linguistic characteristics, with Tulu being particularly underrepresented in computational linguistic research due to its limited annotated datasets. The datasets provided reflect real-world scenarios, including short, informal, and noisy social media posts, accompanied by significant class imbalance across sentiment categories (Hegde et al., 2023).

This study presents a sentiment analysis system leveraging Multilingual BERT (mBERT). The model was fine-tuned on the provided datasets to classify sentiment into various predefined categories. For Tamil, these categories included positive, negative, unknown, mixed_feelings, and neutral, while for Tulu, they comprised positive, negative, not_tulu, mixed, and neutral. Additionally, significant class imbalance was encountered, which was addressed using upsampling techniques.

The system achieved an F1-score of 0.44 for Tamil, securing 6th place, and an F1-score of 0.56 for Tulu, earning 5th place in the competition. These results demonstrate the effectiveness of the approach in tackling sentiment analysis for code-mixed and low-resource Dravidian languages, while also contributing to the development of robust methods for analyzing code-mixed social media text.

2 Related Works

Sentiment analysis plays a vital role in gauging public opinion on social media. Over the years, various approaches have been successful in this domain, ranging from traditional machine learning techniques to more recent deep learning methods. Initially, models like Support Vector Machines (SVMs) and Naive Bayes were widely used for sentiment classification, leveraging handcrafted features such as term frequency and n-grams (Sugitomo et al., 2021). However, the advent of deep learning revolutionized sentiment analysis, with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks improving the ability to capture sequential dependencies in text (Srinivas et al., 2021).

More recently, transformer-based models like BERT and its multilingual variants, such as mBERT, have set new benchmarks in sentiment analysis (Krasitskii et al., 2025). Additionally, stud-

ies on Dravidian languages have highlighted the challenges of sentiment analysis in code-mixed languages, with recent works exploring the use of multilingual models and transfer learning to address these issues (Perera and Caldera, 2024).

Despite advances, challenges remain, especially for underrepresented languages like Tulu, which lack annotated resources. Solutions like data augmentation, upsampling, and fine-tuning transformer models are needed to improve performance on code-mixed, low-resource datasets. This study advances sentiment analysis for Dravidian languages, focusing on Tamil-English and Tulu-English social media text.

3 Methodology

The implementation details and source code are available on github.[1]

3.1 Dataset Description

The datasets provided for this task consisted of train, dev, and test splits. For both Tamil and Tulu, the train and dev datasets had text and category as labels, while the test dataset only contained ID and text as labels.

The Tamil train dataset consisted of 31,122 rows, and the dev dataset consisted of 1,643 rows, distributed as shown in Table 1 and Figure 1 (Chakravarthi et al., 2020).

The Tulu train dataset consisted of 13,301 rows, and the dev dataset consisted of 1,643 rows, as shown in Table 2 and Figure 2 (Hegde et al., 2022).

The test dataset consisted of 3,459 rows for Tamil and 1,479 rows for Tulu.

The datasets were sourced from social media platforms like YouTube, containing real-world, informal, and noisy text. These texts were code-mixed, presenting challenges such as transliteration, spelling variations, and grammatical inconsistencies. Furthermore, significant class imbalance was observed in both Tamil and Tulu datasets, which required careful preprocessing and handling during model training.

3.2 Data Preprocessing

In text classification tasks, effective data preprocessing is crucial to ensure that the model can learn meaningful patterns from the raw data. This involves cleaning, transforming, and structuring raw

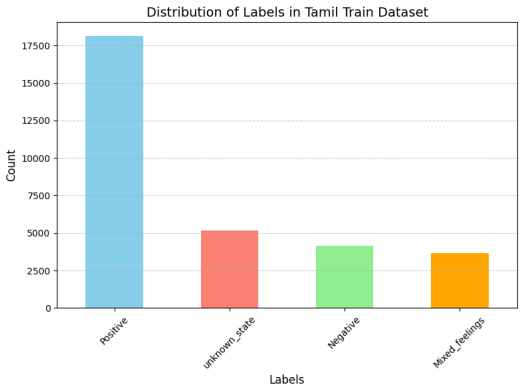


Figure 1: Distribution of Training Labels for Tamil

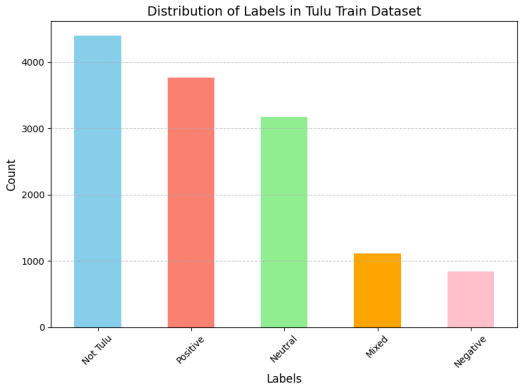


Figure 2: Distribution of Training Labels for Tulu

data into a format suitable for analysis or model training.

One of the key preprocessing steps involved eliminating punctuation marks and special characters. This removal helps reduce noise and inconsistencies in the data, allowing the model to focus on the core content of the text. By minimizing irrelevant elements, the model can better identify the essential features for sentiment analysis.

Tokenization is another crucial preprocessing step in this approach. The input text is tokenized using the BertTokenizer from the pre-trained mBERT model (bert-base-multilingual-cased). The tokenizer splits the text into smaller units, or tokens, that the model can process effectively. These tokens are then padded to a maximum length of 128 and truncated when necessary to fit within the model's input size. This ensures uniformity in input lengths across different samples, allowing for efficient training and model inference.

3.3 Synthetic Sample Generation

A significant challenge encountered during exploratory data analysis was class imbalance, where

[1] https://github.com/DiyaSeshan/DravidianLangTech2025-Sentiment-Analysis/tree/main

	Positive	Negative	Unknown State	Mixed Feelings
Training Data	18,145	4,151	5,164	3,662
Validation Data	2,272	480	619	472

Table 1: Distribution of Sentiment Labels in Tamil Dataset

	Positive	Negative	Neutral	Mixed	Not Tulu
Training Data	3,769	843	3,175	1,114	4,400
Validation Data	470	118	368	143	543

Table 2: Distribution of Sentiment Labels in Tulu Dataset

certain sentiment categories were underrepresented in the dataset. For example, in Tamil, the labels other than Positive were all significantly underrepresented, with Unknown State, Negative, and Mixed Feelings being much less frequent. Similarly, in Tulu, the Mixed and Negative sentiment labels had considerably lower representation compared to the Positive, Neutral, and Not Tulu categories.

To mitigate this issue, RandomOverSampler was used from the unbalanced-learn library. This technique randomly duplicates samples from minority classes, creating a more balanced class distribution. This balances the class distribution, helping to prevent the model from becoming biased towards the majority class, which could negatively impact performance (Permataning Tyas et al., 2023). In this study, RandomOverSampler saturated the number of rows for all labels to match the maximum class size, resulting in 18,145 rows per label in the Tamil training dataset and 4,400 rows per label in the Tulu training dataset.

3.4 Proposed Model

After performing data preprocessing, tokenization, and random oversampling, the dataset was split into training and validation subsets to allow effective evaluation of model performance. Three transformer-based models were experimented with: mBERT, Tamil BERT (from Hugging Face), and IndicBERT (from AI4Bharat). Among the models tested, mBERT demonstrated superior performance due to its multilingual capabilities, making it suitable for both Tamil and Tulu datasets. mBERT is a transformer-based language model pre-trained on 104 languages, making it highly effective for handling code-mixed text and providing robust performance across a wide range of languages, including Tamil and Tulu.

During training, the following hyperparameters were used:

- **Batch size**: Determines the number of training samples used in one forward/backward pass. Set to 16.
- **Maximum token length**: Defines the maximum number of tokens in each input sequence, ensuring the model processes inputs efficiently. Set to 128.
- **Optimizer**: AdamW, which combines the benefits of Adam with weight decay for regularization, often improving generalization.
- **Learning rate**: Controls how much to change the model's weights with respect to the loss gradient during training. Set to 3e-5.

Class weights were computed using compute_class_weight to adjust the loss function. The model was trained for 10 epochs, and evaluation metrics such as accuracy and classification reports were generated to assess performance on the validation set.

While Tamil BERT and IndicBERT also performed reasonably well, they did not match the performance of mBERT on this specific task. Tamil BERT is tailored for the Tamil language but lacks the multilingual capabilities of mBERT. IndicBERT, while promising for Indian languages, likely faced challenges with the code-mixed nature of the dataset. Given these comparisons, mBERT's multilingual training and robustness in handling diverse linguistic inputs made it the optimal choice for sentiment analysis in this study.

4 Results

In the training phase, the model achieved F1-scores of 0.8055 for Tamil and 0.8173 for Tulu, indicating that it successfully captured the sentiment patterns within the training data despite the challenges

posed by the code-mixed nature of the text and the class imbalances present in both languages. Moreover, the model demonstrated strong performance in identifying mixed feelings, unknown states, and neutral sentiments, effectively distinguishing these complex sentiment categories amidst the diverse and noisy social media data.

In the testing phase as well, the model demonstrated notable performance for both Tamil and Tulu. For Tamil, the F1-score achieved was 0.4461, ranking 6th in the ranklist. For Tulu, the model performed slightly better, achieving a F1-score of 0.569, ranking 5th in the ranklist. Despite the challenges posed by Tulu, a low-resource and underrepresented language with limited annotated data and scarce contextual information, the model's performance was commendable. The successful handling of Tulu sentiment analysis showcases the effectiveness of the discussed approach in overcoming barriers such as data sparsity and linguistic underrepresentation.

These results indicate a promising direction for future improvements in sentiment analysis for Dravidian languages.

Metric	Precision	Recall	F1-score
Positive	0.86	0.65	0.74
Negative	0.90	0.94	0.92
unknown_state	0.81	0.94	0.87
Mixed_feelings	0.89	0.92	0.91

Table 3: Performance Metrics for Sentiment Analysis of Tamil

Metric	Precision	Recall	F1-score
Positive	0.77	0.82	0.79
Negative	0.89	0.99	0.93
Neutral	0.87	0.69	0.77
Mixed	0.81	0.93	0.87
Not Tulu	0.90	0.79	0.84

Table 4: Performance Metrics for Sentiment Analysis of Tulu

5 Conclusion

In conclusion, this study presents a successful approach to sentiment analysis for code-mixed Tamil-English and Tulu-English datasets, demonstrating the effectiveness of mBERT in handling the complexities of code-switching and low-resource languages. Despite the challenges posed by noisy social media data and class imbalance, the model achieved promising results, achieving competitive F1-scores across sentiment classes.

The results highlight the model's strength in capturing sentiment nuances, especially for mixed feelings, unknown state, and neutral categories. They reinforce the potential of transformer-based models in advancing sentiment analysis for underrepresented languages and lay the foundation for future improvements through enhanced data preprocessing, augmentation, and fine-tuning strategies.

6 Future Enhancements

While transformer-based models like mBERT have shown promising results in sentiment analysis for code-mixed languages, several areas remain open for improvement. Future research could explore domain-adaptive pretraining, where models are fine-tuned on social media-specific corpora to enhance their understanding of informal and code-mixed text.

Additionally, incorporating external linguistic resources, like code-mixed lexicons and transliteration models, could help improve accuracy, especially for low-resource languages like Tulu. These resources can bridge the gap in language understanding and provide better context for sentiment analysis in code-mixed data.

Finally, integrating multimodal sentiment analysis by combining textual, visual, and audio cues from social media posts could provide a more comprehensive understanding of sentiment, enhancing real-world applications in social media monitoring and customer sentiment analysis.

7 Limitations

The discussed approach relies on a predefined dataset structure, which may limit its generalizability to a variety of real-world scenarios where data distributions differ significantly. Furthermore, the model's computational complexity increases with input length, which reduces its scalability for very long texts without substantial optimization.

Another limitation is the dependence on high-performance hardware, such as GPUs, for efficient training and inference, particularly since the datasets used in this study were extremely large. This could pose challenges for deployment in resource-constrained environments, where access to such computational resources is limited.

References

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, SUBALALITHA CN, Lavanya S K, Thenmozhi D, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Mikhail Krasitskii, Olga Kolesnikova, Liliana Chanona Hernandez, Grigori Sidorov, and Alexander Gelbukh. 2025. Comparative approaches to sentiment analysis using datasets in major european and arabic languages. In *Artificial Intelligence and Big Data Trends 2025*, AIBD, page 137–150. Academy Industry Research Collaboration Center.

Anne Perera and Amitha Caldera. 2024. Sentiment analysis of code-mixed text: A comprehensive review. *JUCS - Journal of Universal Computer Science*, 30(2):242–261.

Salsabila Mazya Permataning Tyas, Riyanarto Sarno, Agus Tri Haryono, and Kelly Rossa Sungkono. 2023. A robustly optimized bert using random oversampling for analyzing imbalanced stock news sentiment data. In *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pages 897–902.

Lavanya S. K., Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, and Rajkumar Charmathi Kumaresan, Prasanna Kumar. 2024. Overview of Second Shared Task on Sentiment Analysis in Code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Akana Srinivas, Ch Satyanarayana, Ch Divakar, and Katikireddy Sirisha. 2021. Sentiment analysis using neural network and lstm. *IOP Conference Series: Materials Science and Engineering*, 1074:012007.

Jason Cornelius Sugitomo, Nathaniel Kevin, Nayra Jannatri, and Derwin Suhartono. 2021. Sentiment analysis using svm and naïve bayes classifiers on restaurant review dataset. In *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, volume 1, pages 100–108.

NLP_goats@DravidianLangTech 2025: Detecting Fake News in Dravidian Languages: A Text Classification Approach

Srihari V K
Sri Sivasubramaniya Nadar College of Engineering
srihari2210434@ssn.edu.in

Vijay Karthick Vaidyanathan
Sri Sivasubramaniya Nadar College of Engineering
vijaykarthick2210930@ssn.edu.in

Thenmozhi Durairaj
Sri Sivasubramaniya Nadar College of Engineering
theni_d@ssn.edu.in

Abstract

The advent and expansion of social media have transformed global communication. Despite its numerous advantages, it has also created an avenue for the rapid spread of fake news, which can impact people's decision-making and judgment. This study explores detecting fake news as part of the DravidianLangTech@NAACL 2025 shared task, focusing on two key tasks. The aim of Task 1 is to classify Malayalam social media posts as either original or fake, and Task 2 categorizes Malayalam-language news articles into four levels of truthfulness: False, Half True, Mostly False and Partly False. We accomplished the tasks using transformer models, e.g., mBERT and classifiers like Naive Bayes. Our results were promising, with mBERT achieving the better results. We achieved a macro-F1 score of 0.83 for distinguishing between fake and original content in Task 1 and a score of 0.54 for classifying news articles in Task 2, ranking us 11 and 4, respectively.

1 Introduction

The rapid growth of social networks has transformed communication, allowing users to express opinions and share content easily. However, this has led to the spread of fake news, defined as false or misleading information presented as real news (Shu et al., 2017). The viral nature of social media amplifies its spread, influencing decisions and eroding trust in genuine sources (Vosoughi et al., 2018), highlighting the need for effective detection mechanisms (Kumar and Shah, 2018).

This challenge is even greater for underrepresented languages such as Dravidian languages, including Malayalam, Tamil, and Telugu, which face a lack of computational resources and annotated datasets for NLP. To address this, DravidianLangTech@NAACL 2025 proposes two tasks: Task 1 classifies the text of social media as fake or original, and Task 2 detects fake news in Malayalam News. These tasks aim to improve fake news detection in Dravidian languages using large datasets and advanced machine learning techniques.

Section 2 reviews related work on fake news detection, especially for Dravidian languages. Section 3 describes the task, and Section 4 outlines the methodology, including details of the data set and the model selection. Section 5 presents experimental results, and Section 6 offers error analysis. Section 7 concludes by summarizing the study's findings.

This study aims to enhance misinformation detection in low-resource languages, focusing on Dravidian languages, by developing effective methods for identifying fake news using models like mBERT and Naive Bayes.

The taken approach is not specific to only detecting fake news, but can also be used to detect abusive language in text towards women as shown in (Rajiakodi et al., 2025). The goal is to contribute to the advancement of NLP and misinformation detection. For implementation, please refer to this GitHub repository (srihari2704).

2 Related Work

The detection of fake news continues today. An evolving field, it has ample scope for improvement, especially in languages with fewer resources and datasets. The improvement in technology and extensive research combined with the availability of datasets for such languages have improved the performance of models to detect fake news in such languages.

The author (Shu et al., 2017) examined methods that integrate content-based characteristics and social context characteristics, highlighting the importance of modeling user behavior and social network analysis. This study opened the door to integrating contextualized information alongside text mining

and motivated selected models that are multimodal and/or hybrid.

A pivotal study, (Wang, 2017), introduced the LIAR dataset. A widely used benchmark for fake news detection, this work used logistic regression and SVM classifiers. It showcased the value of curated datasets. It complements the work by Shu et al., providing the necessary data infrastructure to evaluate approaches combining content and social context.

Using deep learning approaches, (Abualigah et al., 2024) presented the power of neural networks in extracting semantic features for text content. This paper showed how deep learning architectures, particularly Bidirectional LSTMs (BiLSTMs), could generalize better than classical classifiers if combined with the linguistically rich word embeddings GloVe. Building on the earlier focus on features and datasets, this study transitioned the field toward neural approaches for feature extraction and classification.

A multifaceted strategy to counteract Malayalam fake news, (Devika et al., 2024a) extended the work on data set curation by introducing a labeled dataset specifically for Malayalam fake news. Their adoption of multilingual BERT and machine learning classifiers highlighted the challenges of generalizing state-of-the-art techniques to resource-poor environments. This is consistent with using datasets and pre-trained models for language-specific tasks.

The author (Rahman et al., 2024) further emphasized the power of language-specific models, as they obtained the best-shared task F1 score. These studies prove that advanced neural architectures can excel when adapted appropriately.

These studies provide a clear overview of the trajectory of fake news detection. Early research revolved around datasets, feature design, and simple classifiers; further work incorporated deep learning techniques and language-specific approaches. The shift toward such low-resource languages as Malayalam marks a move toward linguistic diversity and tailoring advanced technology to address a global issue.

3 Task Description

The shared task of Fake News Detection in Dravidian Languages aims to address the widespread misinformation on online platforms using given datasets. It is divided into two subtasks:

3.1 Task 1

Classify English social media posts or comments from Twitter, Facebook, and YouTube as *fake* or *original*. The dataset as shown in Figure 1 for this task is provided by previous work on the detection of fake news in Dravidian languages (Subramanian et al., 2025, 2023). Participants develop systems to identify misinformation and ensure the authenticity of online content.

	text	label
1		
2	നല്ല അവതരണ	Fake
3	Masha Allah	Fake
4	അനേഷണം	Fake
5	Illathentha avaru	Fake
6	Barana pakshath	original

Figure 1: Dataset for task 1

3.2 Task 2

Detect and categorize Malayalam news articles into four labels: *False, Half True, Mostly False, Partly False.* The task focuses on addressing misinformation in regional languages to ensure inclusivity and accuracy in local news verification. The dataset as shown in Figure 2 for this task is based on previous studies (Subramanian et al., 2024; Devika et al., 2024b).

FAKE_MAL_TR_0181	അഴീക്കോട് പുലി ഇറങ്ങി.. ര	FALSE
FAKE_MAL_TR_0182	തൃശൂർ ശോഭാസിറ്റി മാളിനു	MOSTLY FALSE
FAKE_MAL_TR_0183	തൃശൂർ ജില്ലാ ആശുപത്രിയിലെ	HALF TRUE
FAKE_MAL_TR_0184	തൃശൂർ പൂരത്തിനു അനുമതി	PARTLY FALSE

Figure 2: Dataset for task 2

4 Methodology

This study designs machine learning models to classify social media comments in the Malayalam language as fake or real, thus countering misinformation. The dataset consisted of labelled comments for supervised learning. Pre-trained multilingual BERT is used to tokenize and process the text for model training. The dataset is split into training and validation sets, and the model is fine-tuned to detect fake news. These performances are reviewed based on accuracy, precision, recall, and F1 score. The goal is to reduce the spread of misinformation and encourage informed discussion in digital communication.

4.1 Data Preprocessing

Preprocessing played a critical role in dataset preparation, enabling better results when predicting fake news. The train and development datasets, each with two columns (text and label), were loaded using Pandas. A quick inspection verified their structure and integrity.

Tables 1 and 2 show that the label distribution was analyzed using Matplotlib bar charts to identify possible class imbalances. To address this, random oversampling was employed to balance the classes in Task 2, as it exhibited a significant imbalance. However, Task 1 was already relatively balanced, so no oversampling was performed. Labels were encoded into binary format, with "Fake" zero and "Original" 1, to ensure consistency and compatibility with machine learning models.

Label	Count
Original	1658
Fake	1599

Table 1: Label distribution of Task 1 dataset.

Label	Count
FALSE	1386
MOSTLY FALSE	295
HALF TRUE	162
PARTLY FALSE	57

Table 2: Label distribution of Task 2 dataset.

Text cleaning removed noise, including punctuation and emojis, using regular expressions to enhance the clarity of the data set. Following cleaning, the text was reviewed to confirm accurate preprocessing.

HuggingFace's AutoTokenizer from the multilingual BERT model (bert-base-multilingual-cased) was used for tokenization, incorporating padding and truncation to 128 tokens for computational efficiency. The datasets were converted into HuggingFace's Dataset format, enabling seamless integration with the training pipeline.

4.2 Model Description

Fake news detection is tackled using various machine learning algorithms and transformer models. Traditional classifiers offer baseline comparisons, while mBERT enhances performance with contextual understanding, allowing evaluation of multiple approaches to identify the most effective solution.

4.2.1 Model selection

Table 3 compares different machine learning and deep learning models for detecting fake news in Malayalam in terms of their performance scores. Among them, M-BERT has the best score (0.85), which shows that it is superior in detecting contextual subtleties in the Malayalam language. Naïve Bayes comes second with a score of 0.80, demonstrating that probabilistic techniques are still influential. Conventional machine learning models such as Logistic Regression, SVM, and MLP demonstrate competitive performance (between 0.77 and 0.79), underscoring their dependability in the face of the increasing popularity of deep learning methods. XLM-R, another transformer-based model, has a score of 0.76, marginally lower than M-BERT but still showing effectiveness. The findings show that transformer-based models such as M-BERT and XLM-R perform better than conventional approaches, supporting contextual embeddings' significance in addressing the Malayalam language's intricacies.

Model	Task1 Score	Task2 Score
Logistic Regression	0.79	0.79
Neural Network (MLP)	0.77	0.77
Support Vector Machines (SVM)	0.78	0.78
MLP (Alternate)	0.72	0.73
Naïve Bayes	0.80	0.80
XLM-R	0.76	0.77
M-BERT	0.85	0.85

Table 3: Performance Comparison of Models for Fake News Detection in Malayalam

4.2.2 mBERT

Multilingual BERT (mBERT) is an extension of BERT designed to manage multilingual inputs. Introduced in (Devlin et al., 2018), it is trained on a corpus that includes 104 languages, enabling cross-language predictions without needing separate models for each language (Devika et al., 2024a). Built on the transformer architecture, mBERT employs self-attention mechanisms to capture contextualized word embeddings across various languages, making it particularly effective for tasks such as fake news detection in Dravidian languages (Wu and Dredze, 2019).

In Task 1, which involves binary classification of social media posts as 'fake' or 'real,' mBERT performs well due to its contextual understanding and ability to generalize across languages.

Table 4 presents the performance report for the mBERT model in Task 1. The model achieved an

accuracy of 0.85, demonstrating its effectiveness in distinguishing between fake and original news.

Metric	Value
Precision	0.8210
Recall	0.8973
F1 Score	0.8575
Accuracy	0.8503

Table 4: Performance Metrics of mBERT for Task 1

For Task 2, where fine-grained classification of Malayalam text is required (such as categorizing posts as "False" or "Half True"), mBERT demonstrates strong performance. Using its multilingual capabilities effectively, it can handle complex linguistic patterns (Pires et al., 2019). Its proficiency in processing low-resource languages highlights its robustness, particularly in data-scarce scenarios (Ruder et al., 2019).

Table 5 displays the mBERT model's performance report in Task 2. The model achieved an accuracy of .80, demonstrating its effectiveness in distinguishing between fake and original news.

Category	Precision	Recall	F1-Score
False	0.79	0.89	0.84
Half True	0.33	0.17	0.23
Mostly False	0.32	0.17	0.22
Partly False	0.08	0.10	0.09

Table 5: Performance Metrics of mBERT for Task 2

mBERT has limitations with underrepresented languages or with significant morphological variations. Fine-tuning for specific tasks, such as fake news detection in Malayalam, improves its performance, making it a valuable tool for multilingual NLP and scalable misinformation detection (Ruder et al., 2019).

5 Results

Among the models used for the detection of fake news, mBERT significantly outperformed the others. Naive Bayes achieved a slightly higher macro F1 Score of 80.23, while mBERT achieved a macro F1 Score of 85.12. This highlights mBERT's ability to effectively represent long-range linguistic structures and context-sensitive elements, particularly for Malayalam, where shallower models proved less effective. The results emphasize the advantages of using advanced transformer-based architectures like mBERT for tasks involving multilingual and morphologically rich languages.

6 Error Analysis

An analysis of the fake news detection task revealed that the Naive Bayes model achieved a macro F1-score of 80.23 but struggled with ambiguous language and overlapping features. In contrast, mBERT performed better, scoring 85.12, although it encountered challenges with rare linguistic constructs and code-mixed content. Both models had difficulty because of the morphological complexity of the Malayalam language and the informal nature of the social media texts. This emphasizes the need for domain-specific pretraining, improved fine-tuning, and exploration of hybrid approaches.

7 Limitations

The primary limitations of this study stem from the challenges associated with processing low-resource languages like Malayalam, which lack large-scale annotated datasets for training robust fake news detection models. Morphological complexity and code-mixed content further hinder model performance, as transformer models like mBERT may struggle with rare linguistic constructs. Additionally, the study relies heavily on pretrained multilingual models, which, while effective, may not fully capture the nuances of Malayalam compared to language-specific models. Another limitation is the difficulty in detecting nuanced misinformation categories, such as subtle satire or partially false claims, which require deeper semantic understanding. Lastly, data imbalance in some categories, particularly in multi-class classification, may have influenced model generalizability, necessitating more refined balancing techniques for improved performance.

8 Conclusion

The mBERT model was identified as the most effective for fake news detection, achieving a macro F1-score of 0.83 in binary classification and 0.54 in multi-class classification. Its ability to capture contextual and semantic nuances in Malayalam text enabled it to outperform Naive Bayes, which struggled with the language's complexities. This underscores the value of advanced language models and emphasizes the need for robust preprocessing and fine-tuning to combat misinformation effectively.

References

Laith Abualigah, Yazan Yehia Al-Ajlouni, Mohammad Sh. Daoud, Maryam Altalhi, and Hazem Migdady. 2024. Fake news detection using recurrent neural network based on bidirectional lstm and glove. *Social Network Analysis and Mining*, 14(1):40.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024a. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024b. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *Proceedings of the 25th International Conference on World Wide Web Companion*, pages 553–558.

Rafael Pires, Eduardo N. Ribeiro, and Luis C. Lamb. 2019. How multilingual is multilingual bert? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pages 5006–5018. Association for Computational Linguistics.

Tanzim Rahman, Abu Raihan, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@DravidianLangTech EACL2024: Fake news classification using Malayalam-BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 223–228, St. Julian's, Malta. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Iryna Gurevych, and Alexander Kementchedjhieva. 2019. A survey of cross-lingual embeddings. *Journal of Artificial Intelligence Research*, 65:405–443.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explorations*, 19(1):22–36.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Y. Wang. 2017. Liar, liar pants on fire: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 422–426.

F. Wu and M. Dredze. 2019. Social media as a sensor of public opinion. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, 2019:1001–1010.

NLP_goats@DravidianLangTech 2025: Towards Safer Social Media: Detecting Abusive Language Directed at Women in Dravidian Languages

Vijay Karthick Vaidyanathan
Sri Sivasubramaniya Nadar Co llege of Engineering
vijaykarthick2210930@ssn.edu.in

Srihari V K
Sri Sivasubramaniya Nadar College of Engineering
srihari2210434@ssn.edu.in

Thenmozhi Durairaj
Sri Sivasubramaniya Nadar College of Engineering
theni_d@ssn.edu.in

Abstract

Social media in the present world is an essential communication platform for information sharing. But their emergence has now led to an increase in the proportion of online abuse, in particular against women in the form of abusive and offensive messages. A reflection of the social inequalities, the importance of detecting abusive language is highlighted by the fact that the usage has a profound psychological and social impact on the victims. This work by DravidianLangTech@NAACL 2025 aims at developing an automated abusive content detection system for women directed towards women on the Tamil and Malayalam platforms, two of the Dravidian languages. Based on a dataset of their YouTube comments about sensitive issues, the study uses multilingual BERT (mBERT) to detect abusive comments versus non-abusive ones. We achieved F1 scores of 0.75 in Tamil and 0.68 in Malayalam, placing us 13 and 9 respectively.

1 Introduction

Social media websites have changed how humans connect, communicate, and socialize. Although such platforms offer several advantages, they have become a haven for abusive online behaviour and abusive behaviour targeting women. Gender-based abuse on the internet consists of insults and abusive and menacing language aimed at degrading, harassing, and silencing women. This type of abuse is due to socially bottled-up resentments and represents serious challenges in fostering a safe online environment. Implications of such texts can be devastating, as many women abandon the platforms because they are subjected to relentless abuse (Jane, 2016). For this reason, identifying and eradicating abusive content is one of the most necessary challenges in developing safe digital environments and the principle of equal opportunity for everybody.

Detecting abusive language in social media is a daunting challenge, even for highly spoken resource languages like English. Abusive content is usually contextual, making it hard to distinguish between actual abusive behaviour and non-abusive comments. The content can be humorous, sarcastic, or satirical, based on the context of the sentence. In addition, abusers commonly use veiled language, slang, and acronyms to disguise what they are doing. Multilingual or cross-lingual abusive detection constitutes a further challenge to the system, as abusive patterns differ depending on language and culture (Waseem et al., 2017). For low-resource languages such as Tamil and Malayalam, these challenges are compounded further by the absence of annotated datasets and linguistic tools. These Dravidian languages, spoken mainly in South India, are culturally rich, have complicated grammar, and employ distinctive scripts, making them computationally challenging. Besides, abusive language in Tamil and Malayalam is hidden in implicit biases, sarcasm, stereotypes, and idiomatic expressions, which need advanced insight and interpretation to capture correctly (Fortuna and Nunes, 2018).

The DravidianLangTech@NAACL 2025 shared task, aside from caring for the current abusive text labelling task itself, also has the broader task of carrying the responsibility to continue the language research work of low-resource languages forward. Tamil and Malayalam have historically been underrepresented in computational linguistics, and this effort is intended to stimulate researchers to apply language analytics to a socially significant problem. The task stimulates creativity and partnerships in content moderation and language processing of underserved languages by curating datasets and systematising the experiment.

In this paper, Section 2 deals with the related work, discussing known detecting abusive languagse, especially in the case of Dravidian languages. Section 3 is an elaborative description of the Task description. Section 4 narrates the methodology adopted in solving the shared tasks. The de-

tails of datasets, preprocessing steps, and models are discussed here. Section 5 details the results of the experimentation. Section 6 details the error analysis. Finally, Section 7 concludes this paper by explaining this research's main findings and contributions.

Focused on improving abusive language detection in low-resource languages, especially Dravidian languages, this study uses models like M-BERT. The research intends to create strong and effective methods for detecting abusive language towards women in less-documented languages. Through this effort, we aspire, to contribute to the field of detection of abusive language and to make the online community a better place for women. For implementation, please refer to this GitHub repository srihari2704

2 Related Work

Abusive content such as hate speech, derogatory language, and cyberbullying represents a considerable challenge for online platforms to provide a safe digital environment, making the detection of abusive language in social media an important area of research. Researchers have devised several strategies to identify such language, from rule-based strategies to sophisticated machine learning and deep learning strategies that can be adapted to a variety of languages and data sets.

(Davidson et al., 2017) proposed an applied machine-learning model to identify hate speech in English tweets by separating it from abusive and non-abusive language. It employed n-gram-based features and logistic regression. This method revealed the role of feature engineering and lexical parsing in recognizing abusive language on social networking services.

With the origin of neural networks, (Badjatiya et al., 2017) proposed deep learning architectures, especially Long-Short-Term Memory (LSTM) networks, to perform hate speech detection. Their study showed significant improvements compared to traditional machine-learning methods by leveraging neural networks' ability to learn meaning and textual contexts. Research on problem-solving with CNNs and related hybrid models was conducted as neural networks developed. (Park et al., 2018) investigated the combination of CNNs with Gated Recurrent Units (GRUs) to detect abusive language, which yielded better results. However, these approaches often struggled with multilingual and low-resource scenarios.

In recent work, transformer-based models have been put into the centre of attention instead of conventional models, and they have been shown to yield better performance in text classification tasks. (Vaswani et al., 2017) described the Transformer architecture, which has served as the basis for various models, including BERT, mBERT, and XLM-RoBERTa. These models use self-attention to capture long-range dependencies in text and are helpful for abusive language detection on multiple datasets.

(Chakravarthi et al., 2021) extended mBERT to detect abusive language in Tamil and Malayalam, drawing attention to the model's capacity to apply to low-resource languages and code-mixed data. XLM-RoBERTa has also helped enhance detection performance, especially in cases involving multilingual environments. (Mozafari et al., 2020) showed its effectiveness in identifying both explicit and implicit abuse in a variety of languages, with an emphasis on complicated linguistic contexts and low-resource settings. Transformation models also have the potential to solve problems related to stereotypes, coded language, and implicit abuse, which makes them prime for contemporary systems of detecting abuses.

These papers describe the evolution of abusive language detection, from simple rule-based, machine learning-based, to deep learning, to transformer-based methods. Initial efforts focused on feature engineering and traditional classifiers, while later research leveraged neural networks for improved contextual understanding. The growing focus on low-resource languages, such as Tamil and Malayalam, emphasizes the increasing importance of linguistic heterogeneity and tailoring powerful models to ensure global safety.

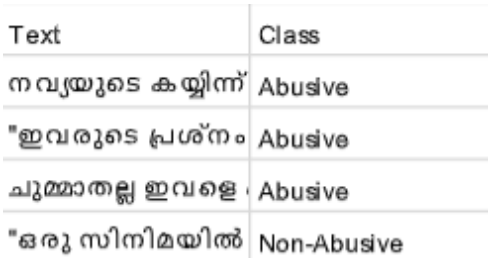
3 Task Description

This study focuses on detecting abusive comments targeting women on social media platforms, specifically in Tamil and Malayalam. The task involves classifying YouTube comments into two categories: Abusive and Non-Abusive. The goal is to identify and address harmful content to promote safer online environments for women. The study aims to improve content moderation by accurately detecting and filtering out abusive language in social media interactions. The dataset for this task is provided by previous works on abusive language

detection in Dravidian languages (Priyadharshini et al., 2023, 2022; Rajiakodi et al., 2025)

Text	Class
இதல்லம் ஒரு தீர்ட	Non-Abusive
யாருடா அந்த கார்;	Non-Abusive
இரண்டு பேரின் (பு	Abusive
என்ன திமிர் இந்த (	Abusive

Figure 1: Tamil Abusive language Dataset



Text	Class
നവ്യയുടെ കയ്യിന്ന്	Abusive
"ഇവരുടെ പ്രശ്നം	Abusive
ചുമ്മാതല്ല ഇവളെ	Abusive
"ഒരു സിനിമയിൽ	Non-Abusive

Figure 2: Malayalam Abusive language Dataset

4 Methodology

The complex nature of Tamil and Malayalam data requires the model to handle linguistic nuances, cultural context, and varied forms of abuse. It must generalize across explicit hate speech, implicit bias, and coded language. The binary classification task demands high precision and recall to accurately identify abusive comments, aiming to enhance content moderation and ensure safer online spaces for women.

4.1 Data Preprocessing

Data preprocessing prepared the raw text for model training, including normalizing Unicode characters, converting text to lowercase, and removing unnecessary punctuation, emojis, and numbers. These steps reduced noise in the data and improved the model's ability to classify abusive comments.

Dataset labels were encoded as numeric values, with "Non-Abusive" as zero and "Abusive" as 1, ensuring consistency for machine learning models.

Overall, these preprocessing techniques improved the quality of the dataset and the model's performance in detecting abusive language.

4.2 Model Evaluation

Recent developments in Natural Language Processing (NLP) have put the strength of transformer-based models in capturing contextual relationships in text irrespective of sequence length into prominence. Among these, XLM-R and mBERT are specially optimized for multilingual applications, including abusive comment classification, allowing models to work even for low-resource languages like Tamil and Malayalam.

Label	Count
Abusive (1)	1531
Non-Abusive (0)	1402

Table 1: Label distribution for Malayalam abusive dataset.

Label	Count
Abusive (1)	1658
Non-Abusive (0)	1598

Table 2: Label distribution for Tamil abusive dataset.

4.2.1 XLM-R

XLM-R (Cross-lingual Language Model - RoBERTa) is a strong multilingual model from the RoBERTa architecture, specially designed to manage multiple languages with cross-lingual pre-training (Conneau et al., 2020). It is highly suited for tasks involving generalization over various language structures, such as abusive comment classification. For our research, XLM-R was fine-tuned to identify abusive comments on social media platforms of Tamil and Malayalam. The model has a transformer encoder that applies special attention mechanisms which enable it to read context forward and backward simultaneously in order to actually upskill in contextual understanding. This bidirectional processing is vital in the detection of subtle patterns of language, particularly in social media posts where context is paramount for the identification of abusive material. The model was trained on preprocessed YouTube comments that had punctuation, emojis, and numbers removed. Tokenization and padding provided uniform input, and the output layer was modified to categorize comments into Abusive and Non-Abusive classes.

4.2.2 mBERT

Multilingual BERT (mBERT), a derivative of BERT, is pre-trained on a large multilingual dataset, thereby allowing it to capture the context of words for many languages (Devlin et al., 2019). This

makes mBERT especially suitable for multilingual text classification tasks such as abusive comment detection. In this study, mBERT was used to classify Tamil and Malayalam social media comments into Abusive or Non-Abusive categories. Similar to XLM-R, mBERT employs a transformer encoder with bidirectional self-attention in order to comprehend left-to-right and right-to-left context equally, which is crucial for identifying abusive material in the subtle nature of social media language. The training data was preprocessed to strip away unnecessary characters, then tokenized and padded. The output layer of mBERT was fine-tuned to categorize comments into two labels: Abusive and Non-Abusive.

Both models proved effective for abusive comment detection, leveraging multilingual pre-training and fine-tuning strategies to perform well even with limited annotated data.

5 Results and Discussion

The mBERT model effectively identified abusive social media comments, recording an F1-score of 0.75 for Tamil and 0.68 for Malayalam. Multilingual pre-training allowed it to understand contextual subtleties, with preprocessing strategies such as noise reduction and tokenization enhancing precision. The model was, however, not good at separating non-abusive Malayalam content, and further fine-tuning or increased datasets would be required.

Another multilingual model, XLM-R, achieved 0.68 on Tamil and 0.65 on Malayalam, falling marginally short of mBERT. Although it also showed excellent multilingual generalization, its slightly lower scores suggest that it would need to be optimized further to perform well at abusive language detection, especially in complicated linguistic contexts.

Overall, mBERT outperformed XLM-R, making it the more suitable model for detecting abusive content in Tamil and Malayalam.

6 Error Analysis

Though mBERT, and XLM-R models demonstrated strong performances in detecting abusive language in Tamil and Malayalam both were held back by some common factors. They frequently misclassified emotionally charged but non-abusive Tamil comments as abusive, indicating sensitivity to certain linguistic patterns. In Malayalam, sarcasm and hidden abuse were regularly missed. They also struggled with code-mixed and transliterated text, reducing its ability to recognize abusive intent accurately. Eg: "Lakshi Ramakrishnan thangalidam our kelvi ketkiren" is classified as abusive even though it is non-abusive.

Addressing these issues requires deeper analysis of false positives, improved handling of sarcasm and implicit abuse, and fine-tuning of model parameters for better classification accuracy.

7 Limitations

Despite achieving promising results, our study has several limitations. The dataset, primarily obtained from YouTube comments, may not fully capture abusive patterns across different social media platforms, including Facebook, Instagram, etc. This may lead to potential bias. The model struggles with sarcasm, implicit biases, and slang, causing occasional false positives and negatives, indicating a need for improved contextual understanding. Limited training data in Tamil and Malayalam impacts the results, while the model's reliance on social media data may hinder its applicability to other domains. Also, the computational load for mBERT and XLM-R is really high when it comes to putting it into real-time servers, especially when we're talking about very lean resources like smaller devices. Plenty of future work should take this on by increasing dataset sizes, incorporating additional knowledge, and refining techniques to get much higher accuracy and fairness.

8 Conclusion

In conclusion, evaluating the mBERT model for abusive comment detection in Tamil and Malayalam highlights its strengths and weaknesses. The ability to process linguistic subtleties contributed to its success in the respective binary classification tasks, with F1-scores reported at 0.75 for Tamil and 0.68 for Malayalam, meaning that the model detected abusive comments reliably in the respective datasets. Similarly, the XLM-R model achieved F1-scores of 0.68 for Tamil and 0.65 for Malayalam, showing slightly lower performance than mBERT.

The model faced issues mainly in differentiating non-abusive content, sometimes causing false positives. These misclassifications indicate a lack of contextual understanding and thus necessitate error mitigation. Future improvements in feature engineering, and fine-tuning can enhance accuracy and robustness, in detecting abusive language.

References

Prakhar Badjatiya, Anupam Gupta, Pavan Kancherla, and Monojit Choudhury. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1406–1411. IEEE.

Bharath Chakravarthi, Harleen Kaur, Ramesh Kumar, and Amit Verma. 2021. Abusive language detection in social media: A survey and new perspectives. In *Proceedings of the 3rd International Workshop on Abusive Language Online (ALW3)*, pages 76–85. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Sebastian Ruder, et al. 2020. Unsupervised cross-lingual representation learning. *arXiv preprint arXiv:2006.03618*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, pages 512–515. Association for the Advancement of Artificial Intelligence (AAAI).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Emma A. Jane. 2016. Online misogyny and feminist digilantism: #mencallmethings, #femfuture, and #solidarityisforwhitewomen. *Continuum: Journal of Media Cultural Studies*, 30(3):284–297.

Mohammad Mozafari, Mohammad Rezaei, Mehdi Ahmadi, Behnam Zeynali, Mohammad Ali Motlagh, Mahmoud Nasiri, and Abolghasem Mahdavi. 2020. Application of machine learning techniques in prediction of human protein–protein interactions: A case study of tuberculosis. *PLOS ONE*, 15(9):e0237861.

Jiho Park, Jihyung Shin, Sangyoun Lee, and Changhyun Seo. 2018. A survey of hate speech detection: Data, methods, and challenges. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 385–395. International Committee on Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 6000–6010. Curran Associates, Inc.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.

HerWILL@DravidianLangTech 2025: Ensemble Approach for Misogyny Detection in Memes Using Pre-trained Text and Vision Transformers

Neelima Monjusha Preeti[1,2], Trina Chakraborty [1,3], Noor Mairukh Khan Arnob[1,4], Saiyara Mahmud[1,4], Azmine Toushik Wasi[1,3]

[1]STEM Team, HerWILL Inc., [2]Jahangirnagar University,
[3]Shahjalal University of Science and Technology, [4]University of Asia Pacific

Correspondence: arnob@uap-bd.edu

Abstract

Misogynistic memes on social media perpetuate gender stereotypes, contribute to harassment, and suppress feminist activism. However, most existing misogyny detection models focus on high-resource languages, leaving a gap in low-resource settings. This work addresses that gap by focusing on misogynistic memes in Tamil and Malayalam, two Dravidian languages with limited resources. We combine computer vision and natural language processing for multi-modal detection, using CLIP embeddings for the vision component and BERT models trained on code-mixed hate speech datasets for the text component. Our results show that this integrated approach effectively captures the unique characteristics of misogynistic memes in these languages, achieving competitive performance with a Macro F1 Score of 0.7800 for the Tamil test set and 0.8748 for the Malayalam test set. These findings highlight the potential of multimodal models and the adaptation of pre-trained models to specific linguistic and cultural contexts, advancing misogyny detection in low-resource settings. Code available at https://github.com/HerWILL-Inc/NAACL-2025

1 Introduction

Misogynistic memes on social media contribute to harmful gender stereotypes and perpetuate inequalities, creating hostile online environments (Chen et al., 2024). These memes amplify sexism, often resulting in online harassment and gender-based cyberbullying (Cai, 2024; Wang and Elfira, 2024). The anonymity and humor inherent in memes offer a unique space to critique societal norms without direct confrontation, providing a platform for both harmful content and progressive movements. While misogynistic memes reinforce oppressive stereotypes, they simultaneously highlight the need for greater awareness, policy intervention, and cultural change. Feminist movements have leveraged

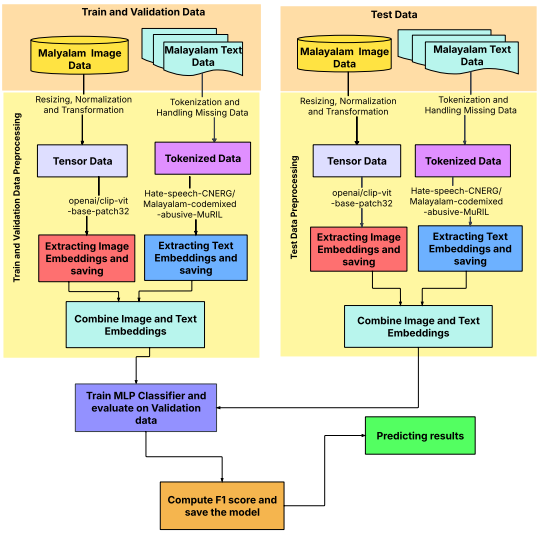


Figure 1: Model architecture, containing tokenizer, pre-trained model, classifier and other components

these platforms to amplify their voices, demonstrating that social media is not only a battleground for gender dynamics but also a powerful space for feminist discourse (Zhu, 2024; Chen et al., 2024). This dual role of memes—simultaneously a tool of oppression and resistance—underscores the complexity of social media's impact on societal norms (Khosravi-Ooryad, 2024).

Misogyny detection is rapidly advancing through the integration of natural language processing (NLP) and computer vision (CV). In NLP, techniques such as classification, severity scoring, and text rewriting help identify harmful language, assess its intensity, and promote respectful discourse. Automated systems on social media platforms effectively detect and flag misogynistic messages, while multilingual models expand detection across languages and cultural contexts (Sheppard et al., 2024; Guzman Cabrera et al., 2024). In CV, models like CLIP (Chen and Chou, 2022) integrate visual and linguistic features to improve detection

accuracy in memes, which combine both text and images. Innovations in image sentiment analysis and graph convolutional networks further enhance the ability to identify misogynistic content. However, there is a significant gap in models for low-resource settings, where the need is greater due to limited datasets and underrepresented languages. Most existing models focus on high-resource languages, leaving a void in addressing gender-based online abuse in these contexts. Although the shared task on Multitask Meme Classification at LT-EDI@EACL 2024 (Chakravarthi et al., 2024) attempted to alleviate this gap, it left room for further improvements.

In this work, we tackled the multi-modal misogyny meme detection task at the Third Shared Task of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025) at NAACL 2025 (Chakravarthi et al., 2025), focusing on Tamil and Malayalam memes. Our approach integrated vision and text modalities to effectively analyze memes, which often blend visual and linguistic elements to convey implicit and explicit misogynistic content. For the vision component, we leveraged CLIP embeddings, a model pre-trained on diverse internet data, to capture the contextual and thematic cues of misogynistic memes. Given CLIP's exposure to web-based imagery, we hypothesized that it would be well-suited for identifying visual patterns associated with misogyny. For the text component, we utilized language-specific BERT models trained on code-mixed hate speech datasets tailored for Tamil and Malayalam. Recognizing that misogynistic memes frequently contain elements of hate speech expressed in code-mixed language, we aimed to capture the linguistic nuances unique to these languages. We combined the embeddings and fine-tuned an MLP classifier, achieving competitive results for both Tamil and Malayalam. Our findings highlight the effectiveness of integrating vision-based CLIP embeddings with language-specific text models for misogyny detection in low-resource languages. This work underscores the importance of multimodal approaches and adapting pre-trained models to specific linguistic and cultural contexts. The implications of this study extend beyond hate speech detection, as it demonstrates the potential for cross-modal learning in tackling social media toxicity, reinforcing the need for AI-driven interventions in promoting safer digital spaces.

2 Problem Description

In this shared task, we were assigned to classify whether a given meme is misogynistic or not. The recently developed MDMD (Misogyny Detection Meme Dataset) dataset (Ponnusamy et al., 2024) was provided for this task. The dataset contains two portions: Malayalam and Tamil. In the Malayalam section, there are 640 memes on the training set, 160 memes on the dev set, and 200 memes on the test set. There are 1136, 284, and 356 memes on the train, dev, and test set of the Tamil part of the dataset. Each meme is recorded as an image file accompanied by textual transcriptions. The shared task was divided into two sub-tasks: Malayalam and Tamil. The training and development set was provided with labels during competition. We submitted our predictions on the test set for the Malayalam sub-task. Solutions were evaluated using macro average F1-score.

3 System Description

Data Pre-processing. In order to guarantee compatibility with pre-trained models and to enable proper embedding extraction, the preprocessing stage involves collecting both textual and visual data. The classification model then uses these embeddings as inputs; as outlined in Figure 1. The procedure includes using the corresponding pre-trained models to handle text and image input separately.

For handling the image data we used the OpenAI CLIP model (`openai/clip-vit-base-patch32`) (Radford et al., 2021) for image embedding extraction. First, Images are resized to 224x224 pixels and normalized into the standard form. Then preprocessed images are transformed into tensors to ensure compatibility with the image encoder model. Finally, extracted embeddings are stored as dictionary with the key-value pair of image-id and embedding tensors.

For handing text data, we employed the pre-trained language model `Hate-speech-CNERG/malayalam-codemixed-abusive-MuRIL` (Das et al., 2022). The transcriptions are tokenized using the corresponding tokenizer. For uniformity, inputs are padded or truncated to a maximum length of 128 tokens. The embeddings are then extracted and stored similarly to image embeddings indexed by image-id. Due to saving image and text embeddings on-disk, we did not have to extract embeddings(which is a time consuming process)

Table 1: Performance in the Validation Set Across Different Models for the Malayalam Dataset

Language Model	Vision Model	F1 Score	Accuracy
ai4bharat/IndicBERTv2-MLM-only	openai/clip-vit-base-patch32	0.8753	0.8812
PosteriorAI/dravida_llama2_7b	zer0int/CLIP-GmP-ViT-L-14	0.8896	0.8938
./malayalam-codemixed-abusive-MuRIL	openai/clip-vit-base-patch32	**0.8940**	**0.9000**

Table 2: Performance in the Test Set Across Different Models for the Tamil Dataset

Language Model	Vision Model	F1 Score	Accuracy
ai4bharat/IndicBERTv2-MLM-only	openai/clip-vit-base-patch32	0.7643	**0.8455**
PosteriorAI/dravida_llama2_7b	zer0int/CLIP-GmP-ViT-L-14	**0.7800**	0.8427
./tamil-codemixed-abusive-MuRIL	openai/clip-vit-base-patch32	0.7575	0.8174

every time we trained our model.

Models. Since misogyny is a form of hate speech, we selected language models pre-trained on offensive corpora to enhance detection performance. For Malayalam, we used Hate-speech-CNERG/malayalam-codemixed-abusive-MuRIL (Das et al., 2022), specifically trained to identify abusive code-mixed Malayalam text. Additionally, we experimented with ai4bharat/IndicBERTv2MLM-only (Doddapaneni et al., 2023), a model trained on 23 Indic languages, including Tamil and Malayalam, to evaluate its generalization capability. To leverage embeddings from a modern LLM, we tested PosteriorAI/dravida_llama2_7b (PosteriorAI, 2024), which has been trained on Kannada, Telugu, Malayalam, and Tamil corpora. For Tamil text encoding, we used Hate-speech-CNERG/tamil-codemixed-abusive-MuRIL (Das et al., 2022), a model specifically designed for offensive Tamil text detection. Given that the dataset consists of memes (internet-based multimodal data), we employed openai/clip-vit-base-patch32 (Radford et al., 2021) as an image encoder, leveraging its training on diverse internet images. To explore potential improvements with a larger model, we also experimented with zer0int/CLIP-GmP-ViT-L-14 (zer0int, 2023). For classification, we trained a lightweight Multi-Layer Perceptron (MLP), ensuring time- and memory-efficient classification while effectively integrating the multimodal embeddings.

Implementation Details. We designed our MLP model for classifying Malayalam memes with an efficient and structured architecture. The model starts with a Linear layer of size $[1280 \times 1024]$, followed by a LeakyReLU activation and Dropout ($p = 0.3$) to mitigate overfitting. Next, we included another Linear layer of size $[1024 \times 512]$, again paired with LeakyReLU and Dropout ($p = 0.3$). Finally, a Linear layer of size $[512 \times 1]$ is followed by a Sigmoid activation function to generate the final classification probability. We set the batch size to 32 and found that training for 10 epochs was sufficient for convergence. We used a learning rate of 0.0005, which provided a good balance between stability and learning speed. Our MLP model had 1,410,323 trainable parameters, making it both lightweight and effective for the task.

4 Experimental Findings

4.1 Malayalam Results

The performance metrics of various models trained on the Malayalam portion of the MDMD dataset are presented in Table 1. The IndicBERT model underperformed due to its limited exposure to offensive and misogynistic language in Malayalam, making it less effective for this task. Similarly, despite its large size (7 billion parameters), the dravida_llama2 model failed to achieve top results, likely because its pre-training corpus lacked sufficient misogynistic text. Our hypothesis that a hate-aware language model would be more effective is strongly supported by the results in Table 1, where the malayalam-codemixed-abusive-MuRIL model achieved the highest F1 score of 0.8940 on the validation set.

Based on this strong validation performance, we submitted our final solution using a combination of predictions from the malayalam-codemixed-abusive-MuRIL (110 million parameters) and the vision-based clip-vit-base-patch32 model. This approach secured **2nd place** in the shared task, achieving an

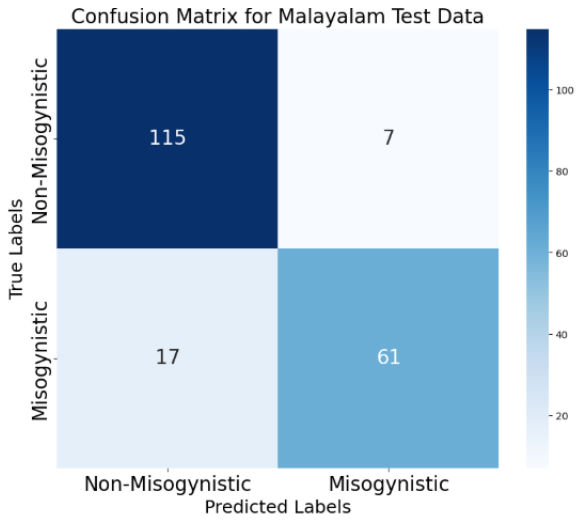


Figure 2: Performance of our proposed model on the malayalam test data

impressive F1 score of 0.8748 on the test set. The confusion matrix in Figure 2 further illustrates the model's effectiveness, correctly classifying a total of 176 out of 200 memes. These results highlight the importance of task-specific pretraining and multimodal integration for improving misogyny detection in low-resource languages.

4.2 Tamil Results

As shown in Table 2, we performed experiments on the released test set with labels to find F1 score and accuracy on the test set. It is clear that the `dravida_llama2` achieves the best F1 score in combination with the large `CLIP-GmP-ViT-L-14` model. The `IndicBERTv2` model achieved the best accuracy of 0.8455 despite being a lightweight model (278 Million Parameters) compared to `dravida_llama2` (7 Billion Parameters). Although the `tamil-codemixed-abusive-MuRIL` model was trained on offensive text, it performed relatively poorly on the test set.

5 Discussion

Our experimental findings highlight several key challenges and insights in misogynistic meme detection for Tamil and Malayalam. One of the primary challenges was achieving robust performance on the Tamil subset of the MDMD dataset. Unlike high-resource languages, where models benefit from extensive labeled datasets, Tamil's low-resource nature limits the effectiveness of pretrained models, leading to underfitting and weaker generalization. Furthermore, our results indicate that increasing model size does not necessarily lead to better performance. The 7-billion-parameter LLaMA2 model, despite its scale, did not outperform smaller transformer-based models fine-tuned on task-specific data. This reinforces the No Free Lunch Theorem (Wolpert, 1996), suggesting that model selection should be guided by domain relevance rather than sheer size. In this task, models explicitly trained on code-mixed and abusive language data demonstrated superior performance over general-purpose large language models. Our findings highlight the value of integrating vision and text for misogyny detection in memes. CLIP embeddings captured contextual cues from images, while fine-tuned BERT models processed code-mixed text. This multimodal approach effectively addressed both implicit and explicit misogynistic content.

Overall, this study highlights the necessity of curating high-quality, domain-specific datasets for low-resource languages and refining model architectures to suit the nuances of code-mixed social media text. Future research should explore adaptive pretraining strategies, knowledge distillation, and cross-lingual transfer learning to enhance performance. Expanding the dataset with more diverse examples and including user interaction patterns could further improve the robustness of misogyny detection systems in Tamil and Malayalam.

6 Conclusion

In this study, we evaluated various models for detecting misogynistic content in Malayalam and Tamil memes. Our results highlight the effectiveness of hate-aware language models, with `malayalam-codemixed-abusive-MuRIL` achieving the highest performance in Malayalam, securing second place in the shared task. This underscores the importance of incorporating offensive text data for low-resource languages. For Tamil, the `dravida_llama2` model combined with `CLIP-GmP-ViT-L-14` yielded the best F1 score, demonstrating the advantages of domain-adapted models. However, models like `IndicBERT` and `tamil-codemixed-abusive-MuRIL` showed mixed results, emphasizing that model size alone is insufficient—training data quality and architecture must be balanced. Our findings contribute to improving misogyny detection in marginalized language communities and lay the groundwork for future advancements in low-resource NLP.

Limitations

One key limitation identified in this study is the absence of diverse offensive and misogynistic words in the available corpora of the Tamil and Malayalam languages. Due to this shortcoming, the models trained on these incomplete corpora perform imperfectly in such an important task as misogynistic meme classification. The CLIP image encoder used in this study was trained on internet data, which is skewed towards developed nations (Radford et al., 2021). Therefore, CLIP lacked the cultural knowledge contained in images of Tamil and Malayalam memes. An image encoder trained on images relevant to Tamil and Malayalam contexts can be developed in the future to achieve better accuracy in misogynistic meme detection. The small dataset size (only 2,776 memes) affects model generalization in this task. Other than curating more data, data augmentation using generative models can be a promising direction to improve the results.

Broader Impact Statement

The findings of this study provide valuable insights for developing systems to detect misogyny in online spaces. Specifically, building a system for detecting misogynistic memes in low-resource languages like Tamil and Malayalam could help reduce the prevalence of misogyny in online communities, particularly those belonging to marginalized groups. However, an important ethical consideration in this area is ensuring the privacy and reputation of individuals depicted in the memes. To mitigate potential harm, we recommend that the research dataset not be publicly released, as doing so could inadvertently perpetuate misogyny. The MDMD dataset should be strictly used for research purposes, with appropriate safeguards in place to protect the privacy and well-being of all stakeholders involved.

References

Ziyan Cai. 2024. The characteristics and causes of the phenomenon of "misogyny" in contemporary chinese online social platforms: Taking weibo and red as examples. *Communications in Humanities Research*, 26(1):113–122.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Lei Chen and Hou Wei Chou. 2022. Rit boston at semeval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from clip model and data-centric ai principle. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Shijing Chen, Usman Naseem, Imran Razzak, and Flora Salim. 2024. Unveiling misogyny memes: A multimodal analysis of modality effects on identification. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1864–1871. ACM.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM conference on hypertext and social media*, pages 32–42.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Rafael Guzman Cabrera, Jose Carmen Morales Castro, Angelica Hernandez Rayas, Jose Ruiz Pinales, and Jose Merced Lozano Garcia. 2024. Automatic detection of misogyny on x using artificial intelligence. *DYNA*, 99(6):562–562.

Sama Khosravi-Ooryad. 2024. Memeing back at misogyny: emerging meme-feminism, visual tactics, and aesthetic world-building on iranian social media. *Feminist Media Studies*, 24(5):984–1003.

Rahul Ponnusamy, Kathiravan Pannerselvam, R Saranya, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, S Bhuvaneswari, Anshid Ka, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality:

Annotated dataset for misogyny detection in tamil and malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488.

PosteriorAI. 2024. Dravida llama 2 7b. Accessed: 2025-01-29.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Brooklyn Sheppard, Anna Richter, Allison Cohen, Elizabeth Smith, Tamara Kneese, Carolyne Pelletier, Ioana Baldini, and Yue Dong. 2024. Biasly: An expert-annotated dataset for subtle misogyny detection and mitigation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 427–452, Bangkok, Thailand. Association for Computational Linguistics.

Ying Wang and Mina Elfira. 2024. *International Review of Humanities Studies*, 9(1).

David H. Wolpert. 1996. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390.

zer0int. 2023. Clip-gmp-vit-l-14. `https://huggingface.co/zer0int/CLIP-GmP-ViT-L-14`. Accessed: 2025-01-29.

Xuanxuan Zhu. 2024. Feminism on social media: Generalized misogyny. *SHS Web of Conferences*, 199:02012.

Cognitext@DravidianLangTech2025: Fake News Classification in Malayalam Using mBERT and LSTM

Shriya Alladi[1], **Bharathi B**[2]
[1] Department of Information Technology
[2] Department of Computer science and Engineering
Sri Sivasubramania Nadar College of Engineering
shriya2310406@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Fake news detection is a crucial task in combating misinformation, particularly in underrepresented languages such as Malayalam. This paper focuses on detecting fake news in Dravidian languages using two tasks: Social Media Text Classification and News Classification. We employ a fine-tuned multilingual BERT (mBERT) model for classifying a given social media text into original or fake and an LSTM-based architecture for accurately detecting and classifying fake news articles in the Malayalam language into different categories.

Extensive preprocessing techniques, such as tokenization and text cleaning, were used to ensure data quality. Our experiments achieved significant accuracy rates and F1-scores. The study's contributions include applying advanced machine learning techniques to the Malayalam language, addressing the lack of research on low-resource languages, and highlighting the challenges of fake news detection in multilingual and code-mixed environments.

1 Introduction

The digital age has amplified the spread of misinformation, posing severe societal and political challenges. While extensive research exists on fake news detection for global languages like English, regional and low-resource languages such as Malayalam remain underexplored. Malayalam, a Dravidian language spoken in southern India, presents unique challenges due to its script, morphology, and prevalence of code-mixed content on social media platforms.

Previous works have shown the efficacy of models such as BERT and LSTM for fake news detection, particularly in multilingual and sequential text processing tasks. However, their application to Dravidian languages remains limited.

This paper addresses this gap by proposing two fake news detection models: an mBERT model for social media text classification and an LSTM-based architecture for classifying news articles. Task 1 involves handling multilingual, code-mixed data with a focus on accurate social media classification, while Task 2 emphasizes sequential dependencies in Malayalam news articles. These contributions aim to enhance fake news detection for low-resource languages and address the societal need to combat misinformation effectively.

The rest of the paper is organized as follows: Section 2 analyzes the related works done in previous research, and Section 3 discusses the dataset. Section 4 contains a detailed discussion of the proposed models used in the current work. Section 5 explains the experimental results. Section 7 highlights the limitations of the study, while Section 7.1 outlines potential future research directions. In Section 6, we conclude the paper. Finally, the Acknowledgment section expresses gratitude to contributors and funding sources.

2 Related Works

Recent research in fake news detection has focused on applying various machine learning algorithms to identify misleading information in online content. (Yuslee and Abdullah, 2021) explore the use of Naive Bayes for fake news detection, highlighting its effectiveness in classifying news articles by leveraging natural language processing (NLP) techniques, including TF-IDF and Count Vectorizer. (Krishna and Adimoolam, 2022) compared the performance of Decision Tree algorithms and Support Vector Machines (SVM) in detecting fake news, emphasizing the reliability and novelty of the Decision Tree approach for fake news detection in social media. Similarly, (Mugdha et al., 2020) evaluate machine learning algorithms, including Naive Bayes, for detecting fake news in Bengali, focusing on feature extraction and classification performance in regional languages. (Ruchansky

et al., 2017) introduce CSI, a hybrid deep model for fake news detection, combining deep learning techniques with traditional methods to improve classification accuracy. Furthermore, (Devlin et al., 2019) present BERT, a deep bidirectional transformer model that has significantly advanced language understanding, and has shown impressive results in text classification tasks, including fake news detection. (Bahad et al., 2019) propose a fake news detection model based on Bi-directional LSTM-Recurrent Neural Network, demonstrating its superiority over other models like CNN, vanilla RNN, and unidirectional LSTM in terms of accuracy for detecting fake news. These studies collectively showcase a variety of machine learning approaches—from classical methods like Naive Bayes and Decision Trees to advanced models like BERT and Bi-directional LSTM—demonstrating their applicability in detecting fake news across different languages and platforms.

3 Dataset Description

he dataset is sourced from the Fake News Detection in Dravidian Languages provided by DravidianLangTech@NAACL 2025 (Subramanian et al., 2024)(Devika et al., 2024)(Subramanian et al., 2023)(Subramanian et al., 2025).

Task 1 consists of a dataset having 4,072 rows, split between the training and validation sets. The training set contains 3,257 articles, while the validation set consists of 815 articles. These articles are labeled as either fake or original, providing a foundation for model training and evaluation in a supervised setting.

Category	Rows
Train Set	3,257
Validation Set	815

Table 1: TASK 1 Data Summary

Task 2 consists of a dataset having a total of 3,120 rows, divided into a training set, validation set, and a test set. The training set for Task 2 contains 1,900 labeled articles, while the validation set holds 200 labeled articles. The test set, crucial for assessing the model's generalizability and performance, consists of 1,020 articles that remain unlabeled, requiring models to predict whether the content is fake or original.

Category	Rows
Train Set	1,900
Validation Set	200
Test Set (Unlabeled)	1,020

Table 2: TASK 2 Data Summary

4 Proposed Methodology

This task involves classifying social media posts in Malayalam as either fake or original news. The process includes several stages: data preprocessing, tokenization, model training, and evaluation. The raw dataset consists of news articles labeled as "Fake" or "Original." Preprocessing involves cleaning the text by removing URLs, mentions, hashtags, and special symbols using regular expressions. Emojis are converted into descriptive text with the emoji.demojize() function, and language detection (via the langdetect library) filters out non-Malayalam or non-English content. Labels are mapped to binary values: "Fake" as 1 and "Original" as 0.

After preprocessing, the text is tokenized using the mBERT tokenizer (bert-base-multilingual-cased), which uses subword tokenization (WordPiece). The sequences are standardized to 128 tokens through padding and truncation. These tokenized sequences, including input IDs and attention masks, serve as input for the model.

The mBERT model is fine-tuned on the preprocessed dataset. The BERT encoder extracts contextual word embeddings, followed by a classification layer with two output neurons (for "Fake" and "Original"). The model is trained using PyTorch with the AdamW optimizer (learning rate = 2e-5) and cross-entropy loss over three epochs. The dataset is split into 80% for training and 20% for validation, with a batch size of 16. The architecture model is present in Figure 1.

The second task involves detecting and classifying fake news in Malayalam articles across multiple categories. Similar to the first task, data preprocessing removes URLs, HTML tags, punctuation, and numbers while converting text to lowercase. The text is tokenized using TensorFlow's Keras Tokenizer, which converts the top 5,000 most frequent words into integer sequences. These sequences are padded to a fixed length of 100 tokens.

The classification model consists of an Embedding layer (100-dimensional vectors), an LSTM layer (128 units) for learning long-range dependen-

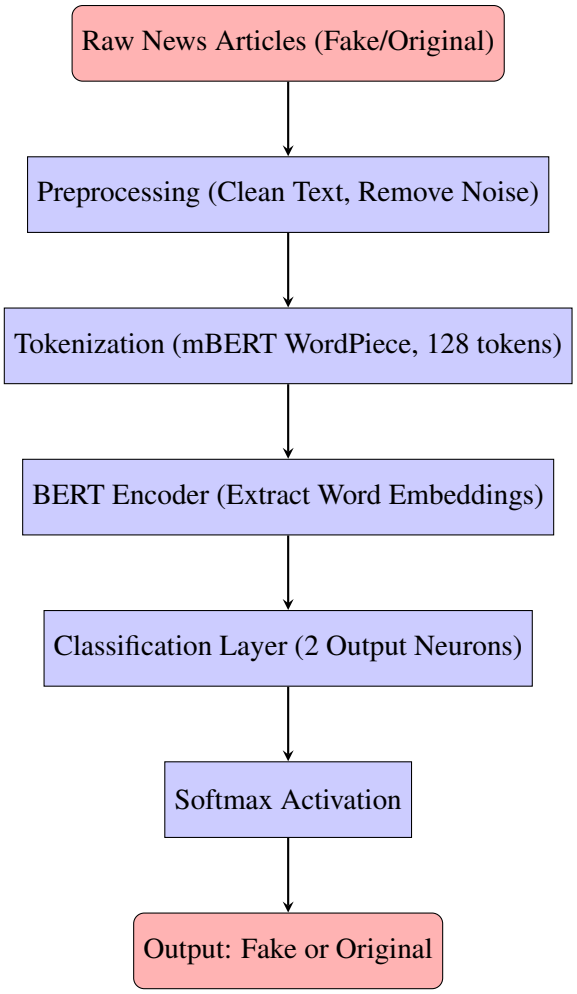


Figure 1: Architecture of the Fake News Detection Model

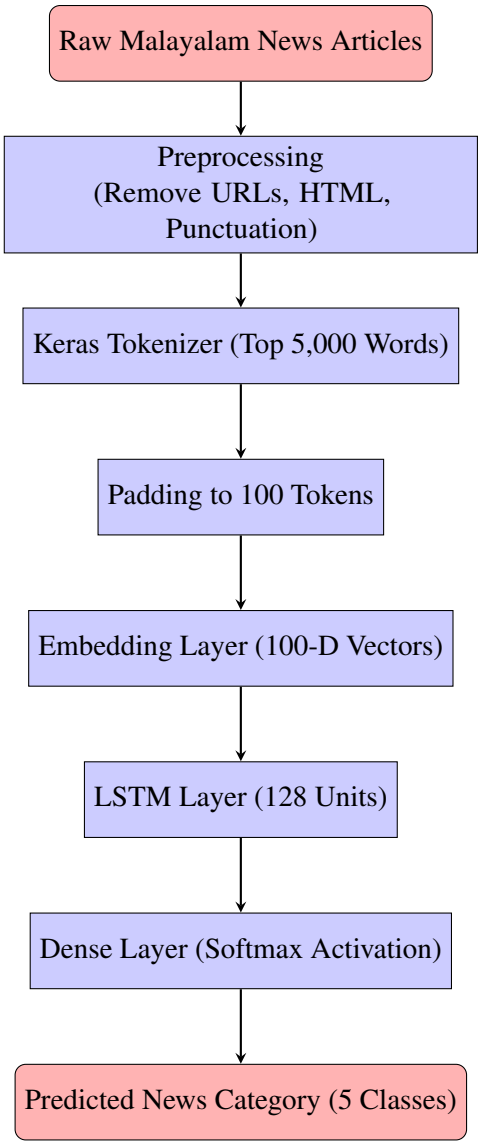


Figure 2: Architecture for LSTM Based Fake News Classification

cies, and a Dense output layer with softmax activation for predicting one of five news categories. The model is compiled with categorical cross-entropy loss, Adam optimizer, and accuracy as the evaluation metric. It is trained with a batch size of 64 for five epochs. The architecture for the same is shown in Figure 2

After training, both models are evaluated on the validation dataset. Performance metrics such as precision, recall, F1-score, and accuracy are computed to assess the effectiveness of the models.

5 Results

This study explores two deep learning approaches—mBERT and LSTM—for fake news classification, demonstrating their effectiveness in identifying deceptive content.

The trained mBERT model was evaluated on the validation dataset, yielding an overall accuracy of 89%, indicating a high level of correctness in classifying news articles. The precision, recall, and F1-score were computed for both classes: "Fake" and "Original." The model performed exceptionally well in identifying original news articles, achieving a precision of 0.94, recall of 0.94, and an F1-score of 0.94. However, the model struggled in identifying fake news, with a precision of 0.50, recall of 0.50, and an F1-score of 0.50. These results suggest that while the model is highly accurate in classifying real news, it has difficulty recognizing fake news articles, possibly due to an imbalance in the dataset, where genuine news articles significantly outnumber fake ones.

The macro average F1-score of 0.72 highlights this disparity, indicating that the model does not generalize equally across both classes. The

Class	Precision	Recall	F1-score
Original	0.94	0.94	0.94
Fake	0.50	0.50	0.50
Accuracy		0.89	
Macro Avg	0.72	0.72	0.72
Weighted Avg	0.89	0.89	0.89

Table 3: mBert Classification Report Results

weighted average F1-score remains 0.89, reinforcing the idea that the model's performance is heavily skewed towards accurately classifying "Original" news, while its capability to detect fake news remains suboptimal. A likely cause for this imbalance is the insufficient representation of fake news samples in the dataset, which could have led to the model learning patterns that favor the majority class.

Class	Precision	Recall	F1-score
Original	1.00	1.00	1.00
Accuracy		1.00	
Macro Avg	1.00	1.00	1.00
Weighted Avg	1.00	1.00	1.00

Table 4: LSTM Classification Report Results

This study explores the use of mBERT and LSTM networks to enhance the accuracy of fake news classification specifically for Malayalam.[1]

6 Conclusions

An effective deep learning approach for fake news classification using an LSTM-based model, demonstrating strong performance in identifying deceptive content. By implementing a robust text preprocessing pipeline and leveraging word embeddings, the model successfully captures contextual nuances, leading to high classification accuracy. The near-perfect performance on the test set suggests that the model has learned meaningful patterns; however, the possibility of overfitting necessitates further investigation. Future work can focus on enhancing generalization through techniques such as dropout regularization, fine-tuning transformer-based architectures like BERT, and expanding the dataset to include diverse linguistic variations. This research underscores the potential of deep learning in combating misinformation and lays the groundwork for more sophisticated models capable of real-world deployment.

To further advance fake news detection, future research can explore innovative data augmentation techniques, integrate valuable metadata features, and experiment with cutting-edge transformer architectures like XLM-Roberta to enhance multilingual text classification. Additionally, addressing class imbalance through methods such as oversampling, class weighting, and dropout regularization will contribute to improved model performance and robustness. This study highlights the promising potential of deep learning in combating misinformation and paves the way for developing more powerful models, offering exciting opportunities for real-world deployment and impactful solutions.

7 Limitations

Despite achieving promising results, our proposed methodology has certain limitations. One major limitation is the class imbalance in the dataset. This imbalance likely contributed to the lower precision and recall scores for fake news detection, as the model struggled to learn representative patterns for the minority class. Additionally, while mBERT effectively captures contextual information, it may not fully account for nuanced linguistic characteristics in Malayalam, especially in code-mixed and informal social media texts. The reliance on subword tokenization can also lead to fragmented representations of rare or morphologically complex words, affecting classification accuracy. Moreover, the LSTM-based approach for news article classification, although effective, may not generalize well to unseen data due to overfitting risks associated with limited training samples. During training, it achieved 100% accuracy, which is indicative of overfitting. This suggests that the model has learned patterns specific to the training data rather than generalizing well to unseen samples. Lastly, our approach does not incorporate external knowledge sources, such as fact-checking databases, which could enhance the model's ability to verify news credibility beyond textual patterns alone. Addressing these limitations in future work could lead to more robust and reliable fake news detection models.

7.1 Future Work

There are several ways to improve our fake news detection model in the future. First, we can ad-

[1] https://github.com/ShriyaAl/Fake-News-Detection-DravidianLangTech2025

dress the class imbalance by adding more fake news samples using data augmentation techniques like back-translation or synthetic text generation. Using fact-checking databases or real-time news verification APIs could also help improve accuracy by providing additional context.

We can explore other transformer models like XLM-Roberta or IndicBERT, which might work better for Malayalam and other low-resource languages. Training these models on a larger dataset with more diverse sources, such as blogs and user comments, could make them more effective at handling different writing styles.

Another important step is improving model interpretability by using attention maps or explainability techniques like SHAP, which help understand how the model makes decisions.

Acknowledgment

We sincerely thank the organizers of DravidianLangTech-2025 at NAACL 2025 for providing the datasets and valuable guidance for this shared task. https://sites.google.com/view/dravidianlangtech-2025/shared-tasks-2025

References

Pritika Bahad, Preeti Saxena, and Raj Kamal. 2019. Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Computer Science*, 165:74–82. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC - DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

N. Leela Siva Rama Krishna and M. Adimoolam. 2022. Fake news detection system using decision tree algorithm and compare textual property with support vector machine algorithm. In *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, pages 1–6.

Shafayat Bin Shabbir Mugdha, Sayeda Muntaha Ferdous, and Ahmed Fahmin. 2020. Evaluating machine learning algorithms for bengali fake news detection. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 797–806. ACM.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Nurshaheeda Shazleen Yuslee and Nur Atiqah Sia Abdullah. 2021. Fake news detection using naive bayes. In *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET)*, pages 112–117.

NLP_goats@DravidianLangTech 2025: Detecting AI-Written Reviews for Consumer Trust

Srihari V K
Sri Sivasubramaniya Nadar College of Engineering
srihari2210434@ssn.edu.in

Vijay Karthick Vaidyanathan
Sri Sivasubramaniya Nadar College of Engineering
vijaykarthick2210930@ssn.edu.in

Mugilkrishna D U
Sri Sivasubramaniya Nadar College of Engineering
mugilkrishna2210314@ssn.edu.in

Durairaj Thenmozhi
Sri Sivasubramaniya Nadar College of Engineering
theni_d@ssn.edu.in

Abstract

The rise of AI-generated content has introduced challenges in distinguishing machine-generated text from human-written text, particularly in low-resource languages. Identifying artificial intelligence (AI)-based reviews is important to preserve trust and authenticity on online platforms. The Shared Task on Detecting AI-Generated Product Reviews in Dravidian languages deals with detecting AI-generated and human-written reviews in Tamil and Malayalam. To solve this problem, we specifically fine-tuned mBERT for binary classification. Our system achieved 10th place in Tamil with a macro F1-score of 0.90 and 28th place in Malayalam with a macro F1-score of 0.68, as the NAACL 2025 organizers reported. The findings demonstrate the complexity of separating AI-derived text from human-authored writing, with a call for continued advances in detection methods. The fine-tuned mBERT model achieved high performance for Tamil, macro F1-score of 0.90 and a score of 0.68 Malayalam. This highlights that some inherent challenges still persist in processing low-resource languages and further language-specific enhancements are needed.

1 Introduction

E-commerce has transformed consumer behaviour, enabling them to share product experiences through reviews on platforms like Amazon and Flipkart. However, the increasing use of AI-generated reviews raises concerns about authenticity, trust, and misinformation in digital markets (Li et al., 2022). AI-powered reviews can manipulate ratings, deceive consumers, and undermine trust, making distinguishing between human- and machine-generated content difficult. Advances in AI text generation further exacerbate this issue, necessitating effective detection mechanisms (Zellers et al., 2019).

Detecting AI-generated reviews is challenging across languages due to the sophistication of generative models and the scarcity of high-quality labelled data. The problem is even more severe in underrepresented Dravidian languages like Tamil, Malayalam, and Telugu, which lack sufficient computational resources and annotated datasets. Their complex linguistic structures, deep morphology, and code-mixing further complicate detection. Reliable AI detection strategies are crucial to maintaining trust in online marketplaces. This study also opens the door for future work, which might explore alternative architectures and larger datasets to overcome current limitations.

This shared task addresses these challenges with two subtasks: Task 1 differentiates human-written reviews from AI-generated reviews in a given dataset, while Task 2 identifies AI-written reviews particularly in Tamil and Malayalam.

This paper is structured as follows: Section 2 reviews prior work on AI-based text detection in low-resource languages, Section 3 details the task descriptions, and Section 4 outlines the methodology, including data preprocessing and model selection and additional model implementations. Section 5 presents experimental results, followed by error analysis in Section 6. Finally, Section 7 concludes with key findings and contributions.

Our study focuses on improving AI-generated review detection in low-resource languages. We aim to develop reliable methods for identifying fake reviews using models like M-BERT, XLM-R and classifiers like Naïve Bayes, contributing to advancements in NLP and AI-generated content detection.

For implementation, please refer to this GitHub repository (srihari2704).

2 Related Work

Detecting AI-generated text is still an open problem because models such as GPT-4 and ChatGPT

generate increasingly human-like text (Brown et al., 2020). Though detection based on linguistic heuristics and statistical approaches once dominated, nowadays, deep learning and transformers are preferred. It is incredibly challenging in product review contexts, where AI-based contents replicate human styles, requiring further effort for detection (Ippolito et al., 2020). With the increasing popularity of AI-aided review generation, efficient detection methods are essential to guarantee the genuineness of online platforms (Zhang et al., 2020).

In (Fagni et al., 2021), the author investigated AI-produced fabricated content in online reviews, presenting a dataset TweepFake, which includes human-natively and AI-infused product-based and social media reviews. They compared sentiment, coherence, and repetition between AI and human written reviews. Their experiments demonstrated that optimized transformer-based models (e.g., BERT, Roberta) helped traditional model classifiers by using attention-based mechanisms to identify inconsistencies in syntactic patterns between AI-generated reviews. The study concluded that detection accuracy improves significantly when classifiers are trained on domain-specific AI-generated review data rather than general-purpose datasets. Work in artificial product review generation has also yielded clues to enhancing detection.

(Li et al., 2020) studied methods for generating deceptive reviews using AI models and analyzed their effectiveness in fooling human evaluators. Their findings demonstrated that state-of-the-art generative models could produce realistic yet generic-sounding reviews, often lacking the nuanced storytelling in human-authored content. These findings indicate that detecting AI-generated reviews should target linguistic patterns, i.e., illogical coherence, redundant sentences, and high sentiment repetition.

The paper (Wu et al., 2021) also investigated the creation and recognition of AI-generated reviews and their implications for e-commerce synthetic content. Their study examined how AI-generated reviews impact consumer trust and purchasing decisions, underscoring the need for detection frameworks that incorporate both linguistic and behavioural features. Their study suggested hybrid models that integrate BERT-based classification with user behaviour analysis and found that allowing the use of metadata–that is, metadata about review times and user actions–to influence the detection system could significantly improve

ID	DATA	LABEL
TAM_HUAI_TR_001	இந்த சோப்பின் ம	AI
TAM_HUAI_TR_002	தோலை நன்கு சுத்	AI
TAM_HUAI_TR_003	இதைப் பயன்படுத்	AI
TAM_HUAI_TR_004	இந்த சோப்பில் இ	AI
TAM_HUAI_TR_005	சிறிது சோப்பு போ	AI

Figure 1: Dataset for Tamil

MAL_HUAI_TR_398	പേരാമംഗലംരാമ	HUMAN
MAL_HUAI_TR_399	കപ്പയും , മീൻ ക	HUMAN
MAL_HUAI_TR_400	നന്നായിട്ടുണ്ട്. ചോ	HUMAN
MAL_HUAI_TR_401	ഞാൻ ഈ ഫോസ്	AI

Figure 2: Dataset for Malayalam

detection performance.

Previous works have demonstrated the effectiveness of transformer models such as mBERT in cross-lingual tasks. However, alternative models like XLM-R have great potential, given low-resource settings, acting as an important direction for future comparisons.

In addition to transformer-based approaches, models such as Logistic Regression have also been utilized. Although these studies have contributed immensely to the detection of AI-written product reviews, nothing is available in Dravidian languages such as Tamil and Malayalam. These languages have high morphological complexity, which makes adapting to new detection models somewhat difficult. The present study seeks to close this gap by training tailor-made transformer-based models for AI-driven review detection against Dravidian languages to provide more stable e-commerce and digital platforms.

3 Task Description

The task aims to detect AI-generated product reviews in Dravidian languages like Malayalam and Tamil, ensuring authenticity for consumer trust. Participants develop models to distinguish human reviews from AI-generated reviews, using data sets from previous studies (Premjith et al., 2025). Figures 1 and 2 show the Tamil and Malayalam datasets.

4 Methodology

Classifying AI-generated and human-written reviews in Tamil and Malayalam is challenging due to their complex linguistic structures. The model

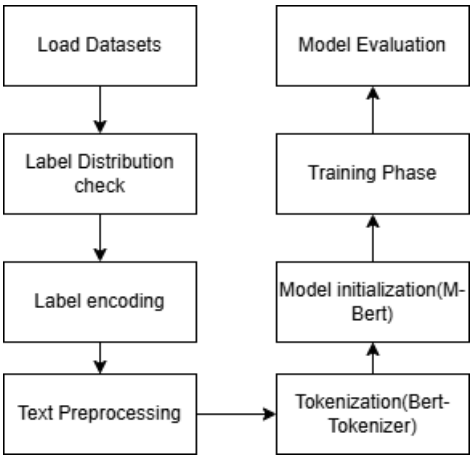


Figure 3: Flowchart representing the process of detecting AI-generated vs human-written reviews

must capture linguistic variation, context, and style for accurate binary classification. The goal is to ensure content authenticity and enhance the reliability of online reviews in Dravidian languages.

Figure 3 represents the overall process for classifying the AI-generated product review.

4.1 Data Preprocessing

Effective preprocessing is essential for improving model performance and distinguishing AI-generated from human-written product reviews in Tamil and Malayalam. The raw dataset undergoes several preprocessing steps to clean and standardize the text.

First, we address missing and inconsistent data by replacing missing text entries with an empty string and mapping non-standardized labels to "AI" for machine-generated reviews and "HUMAN" for human-written reviews.

Next, we clean the text by removing special characters, punctuation, unrelated symbols, non-Tamil/Malayalam symbols, and numerical digits, ensuring linguistic consistency. The text is then converted to lowercase, and redundant spaces are normalized.

After cleaning, the text is tokenized using a BERT-based multilingual tokenizer, resulting in sequences of subword tokens. Shorter sequences are padded to maintain a fixed input length of 256 tokens, while longer sequences are truncated.

Label encoding converts the categorical labels into numerical values, assigning 0 to "AI" and 1 to "HUMAN" for effective processing in a supervised learning context.

The dataset is further stratified into training (80) and testing (20) subsets. This way, AI-generated and human-written reviews are proportionally represented in each subgroup, avoiding class imbalance problems. The label distribution for both AI-generated and human-written reviews in the Tamil and Malayalam datasets is presented in Table 2 and Table 1, respectively. These preprocessing techniques help optimize the dataset to train a robust classification model that can distinguish between AI-generated and human-authored product reviews in Dravidian languages.

Label	Count
AI	405
HUMAN	403

Table 1: Label Distribution in the Malayalam Dataset

Label	Count
AI	410
HUMAN	398

Table 2: Label Distribution in the Tamil Dataset

4.2 Model Evaluation

Recent advancements in Natural Language Processing (NLP) have demonstrated the remarkable capabilities of transformer-based models, especially in tasks involving cross-linguistic text classification. Among these models, mBERT, and XLM-R have shown significant promise in capturing complex contextual information across languages, making them highly effective for text classification tasks such as detecting AI-generated product reviews.

The mBERT model has been pre-trained on a diverse, multilingual corpus, including Dravidian languages like Tamil and Malayalam, which are often considered low-resource languages in the context of NLP (Pires et al., 2019). This extensive pre-training enables mBERT to handle many linguistic features and language structures, which results in a strong performance on tasks with limited annotated data.

XLM-R (Cross-lingual Language Model - RoBERTa) is a strong multilingual model from the RoBERTa architecture, specially designed to manage multiple languages with cross-lingual pre-training (Conneau et al., 2020). It is highly suited for tasks involving generalization over various language structures, such as abusive comment classification.

Also, a Logistic Regression classifier was utilized with TF-IDF features from the preprocessed

text. This method exploits statistical patterns within the text in which word frequency and significance are employed to predict the reviews as being either human or AI-written. Although being straightforward relative to more complex models such as M-BERT and XLM-R, Logistic Regression is a benchmark to gauge the effect of feature engineering and offers a useful benchmark for low-resource scenarios.

5 Results and Discussion

We experimented with the performance of M-BERT, XLM-R, and Logistic Regression models in detecting AI-generated product reviews in Tamil and Malayalam.

M-BERT performed excellently with 0.94 precision, 0.95 recall, and 0.94 F1-scores for Tamil AI class, and 0.95, 0.94, and 0.94 respectively for HUMAN. For Malayalam, M-BERT obtained 0.91 precision for AI and 0.93 for HUMAN with respective F1-scores.

XLM-R marginally improved over M-BERT in Tamil recall (precision 0.95, recall 0.96, F1 0.92) and was considerably better for Malayalam (F1 0.75 vs. 0.68 for M-BERT), showing that it manages Malayalam's language intricacies more effectively.

Logistic Regression, with a general accuracy of 0.82, fared poorer than both M-BERT and XLM-R. Its F1-scores were 0.83 for Tamil AI, 0.81 for Tamil HUMAN, 0.83 for Malayalam AI, and 0.81 for Malayalam HUMAN, revealing the weakness of this model, particularly in dealing with Dravidian language intricacies.

Overall, both M-BERT and XLM-R performed better than Logistic Regression, with the latter being notably better for Malayalam. This indicates that transformer-based models such as M-BERT and XLM-R perform better in AI-generated review detection for Tamil and Malayalam languages, with the latter being the best fit for Malayalam.

6 Error Analysis

The mBERT model performed well in detecting AI-generated reviews in Tamil and Malayalam but faced challenges in misclassifying specific human-written reviews, especially in Tamil. This was due to linguistic features resembling those of AI-generated content. Similar misclassifications were observed in Malayalam, indicating the model's difficulty in capturing subtle contextual cues.

Despite a balanced dataset, errors, primarily false positives, highlight issues in classifying AI-generated content in low-resource languages. Fine-tuning the model's parameters is necessary to improve accuracy and make the system more reliable in classifying reviews correctly.

7 Limitations

The research is confronted with a number of limitations, mostly because of the difficulties of low-resource languages such as Tamil and Malayalam, which do not have enough annotated datasets and computational resources for successful AI-generated text detection. The poorer performance in Malayalam (macro F1-score of 0.68) as opposed to Tamil (macro F1-score of 0.90) reflects the challenge of detecting linguistic subtleties, morphology, and code-mixed forms. The use of transformer-based models like mBERT and XLM-R, although useful, is still prone to missing subtle contextual signals, and therefore misclassifies text, especially in human-composed reviews which are written in a similar AI-like style. Furthermore, the work concentrates on binary classification and does not address more complex cases, for example, mixed content of AI and human. Future research may overcome these limitations by using bigger datasets, more sophisticated fine-tuning methods, and ensemble models that combine linguistic and behavioral features for better detection performance.

8 Conclusion

Finally, a comparison of the mBERT model for detecting reviews generated using AI in Tamil and Malayalam demonstrates its advantages and limitations. The capacity for language nuance captured by it accounted for its effectiveness, obtaining F1-scores of 0.90 in Tamil and 0.68 in Malayalam, suggesting accurate performance for binary classification. Analogously, the XLM-R model also attained 0.92 for Tamil and 0.75 for Malayalam, demonstrating more competent management of language complexities.

Yet, the model was plagued with misclassifications, especially in human-like AI text, resulting in false positives. The glitches indicate shortcomings in contextual understanding, calling for enhancements. Future research should prioritize enhanced fine-tuning, data augmentation, and hybrid strategies to improve detection precision in varied linguistic contexts.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Alexis Conneau, Guillaume Lample, Sebastian Ruder, et al. 2020. Unsupervised cross-lingual representation learning. *arXiv preprint arXiv:2006.03618*.

Tommaso Fagni, Fabrizio Falchi, et al. 2021. Tweepfake: About detecting deepfake tweets. In *Proceedings of the 43rd European Conference on Information Retrieval*, pages 225–238.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, et al. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822.

Jiwei Li, Will Wang, et al. 2020. Fooling humans with ai-generated reviews: An analysis. *Journal of Artificial Intelligence Research*, 69:125–147.

Jiwei Li, Xinyuan Zhang, et al. 2022. Ai-generated reviews and their impact on online marketplaces. *Journal of Artificial Intelligence Research*, 75:1123–1145.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Lingfei Wu, Xinyuan Zhang, et al. 2021. Mind the fake review: Implications of ai-generated reviews for e-commerce. *ACM Transactions on the Web*, 15(4):1–26.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, et al. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9054–9065.

Ying Zhang et al. 2020. Overview of fake review detection methods: From heuristic rules to deep learning. *IEEE Transactions on Computational Social Systems*, 7(5):1236–1248.

RATHAN@DravidianLangTech 2025: Annaparavai - Separate the Authentic Human Reviews from AI-generated one

Jubeerathan Thevakumar
Dept. of Computer Sci. and Eng
University of Moratuwa
Colombo, Sri Lanka
jubeerathan.20@cse.mrt.ac.lk

Luheerathan Thevakumar
Jaffna, Sri Lanka
the.luheerathan@gmail.com

Abstract

Detecting AI-generated reviews is crucial for maintaining the authenticity of online feedback in low-resource languages like Tamil and Malayalam. We propose a transfer learning-based approach using embeddings from XLM-RoBERTa, IndicBERT, mT5, and Sentence-BERT, validated with five-fold cross-validation via XGBoost. These embeddings are used to train deep neural networks (DNNs), refined through a weighted ensemble model. Our method achieves 90% f1-score for Malayalam and 73% for Tamil, demonstrating the effectiveness of transfer learning and ensembling for review detection. The source code is publicly available to support further research and improve online review systems in multilingual settings.

1 Introduction

As artificial intelligence technologies become increasingly sophisticated, the proliferation of AI-generated reviews presents a growing threat to the integrity of online consumer feedback systems. Recent studies have revealed that a significant portion of reviews in sectors such as home services, legal, and medical fields are likely fraudulent, with many confirmed as AI-generated (Karaş, 2024; Thilagavathi et al., 2024). These fake reviews undermine consumer trust, create unfair competition, and pose significant challenges for e-commerce platforms and consumers alike. The rapid production of convincing fake reviews threatens the foundational trust mechanism of online marketplaces, necessitating robust detection systems and enhanced consumer protection measures to maintain the integrity of online review ecosystems.

This research focuses on detecting AI-generated product reviews in Tamil and Malayalam, two low-resource languages spoken in South India. The increasing presence of fraudulent online reviews in these languages underscores the urgent need for effective detection methods. However, the scarcity of linguistic resources and tools for these low-resource languages presents significant challenges. To mitigate these limitations, we utilize two datasets introduced by (Premjith et al., 2025), which comprises both AI-generated and human-authored product reviews in Tamil and Malayalam.

We employed a transfer learning-based approach for feature extraction, utilizing embeddings from four different models. These embeddings are evaluated through cross-validation using XGBoost to assess their discriminative capacity. Following this, we train four independent deep neural network (DNN) models on the extracted embeddings. Finally, we construct an ensemble model that aggregates predictions from the individual models, aiming to improve classification performance through weighted averaging. Our implementation is publicly available in the GitHub[1] repository.

The findings from this study have important implications for strengthening content moderation systems in e-commerce platforms, ultimately fostering greater transparency and trust in online review ecosystems.

2 Related Work

The task of detecting AI-generated product reviews is a subset of the broader AI-generated text detection challenge. While most research in this area has focused on widely spoken languages, there is a notable lack of studies addressing AI-generated content in Tamil and Malayalam.

[1] https://github.com/Jubeerathan/Annaparavai

(Ippolito et al., 2019) employed a set of BERT-based classifiers (Devlin et al., 2019) with three popular random decoding strategies—top-k, nucleus, and temperature sampling—on text samples generated by GPT-2 (Radford et al., 2019). (Fagni et al., 2021) introduced a set of sequence-based classifiers, including LSTM, GRU, and CNN, for detecting AI-generated social media texts.

RoBERTa, a pretrained, non-generative language model (Liu, 2019), was integrated into classifiers to detect text generated by GPT-2 (Solaiman et al., 2019). Despite having a distinct architecture and tokenizer compared to GPT-2, the RoBERTa-based classifier was able to detect text generated by the GPT-2 model with an accuracy of approximately 95%.

Stylometric features, which are quantitative characteristics of a person's writing style, can be used alongside pre-trained language models to enhance detection capabilities. These features highlight the stylistic differences between human and AI authors, aiding in the detection of AI-generated text. Incorporating stylometric aspects such as phraseology, punctuation, and linguistic diversity into pre-trained language model-based classifiers has shown improved performance in detecting AI-generated tweets (Kumarage et al., 2023). Ensemble learning techniques, combined with stylometric features, Linguistic Word Inquiry, GPT-2 word embeddings, and Author's Multilevel Ngram Profiles (AMNP) features, are utilized alongside transfer learning (Mikros et al., 2023) to identify the AI-generated text.

Similar to stylometric features, other notable efforts have focused on leveraging various text characteristics to enhance detection capabilities. SeqXGPT, for example, uses sentence-wise log probability metrics from white-box LLMs to identify AI-generated text at the sentence level (Wang et al., 2023). GPT-who revisits the Uniform Information Density (UID) hypothesis, proposing that AI-generated text may lack the evenness in information distribution typical of human language, and introduces UID features to measure the smoothness of token distribution (Venkatraman et al., 2023). Additionally, another approach improves detection accuracy by combining the factual structure of text with a RoBERTa-based classifier (Zhong et al., 2020). These methods utilize structural and sequential features to enhance the detection of AI-generated content.

Most of these studies collectively underscore the critical role of transformer-based architectures in addressing the challenges of detecting AI-generated content, especially in the English language. By refining language-specific models and exploring multimodal techniques, these research efforts have created a solid groundwork for future progress in the field of AI-generated content detection.

3 Dataset

We used two data sets for this investigation: the Tamil and Malayalam datasets from (Premjith et al., 2025). The Tamil dataset consists of 808 samples in the training set, and 100 samples in the given test set. The Malayalam dataset contains 800 samples in the training set, and 200 samples in the given test set. Both training datasets are annotated with labels Human and AI.

Figures 1 and 2 illustrate the length distribution of the training datasets and testing datasets of each language.

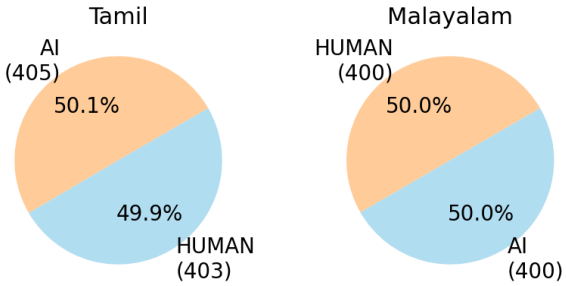


Figure 1: Distribution of labels in Tamil and Malayalam languages in train dataset.

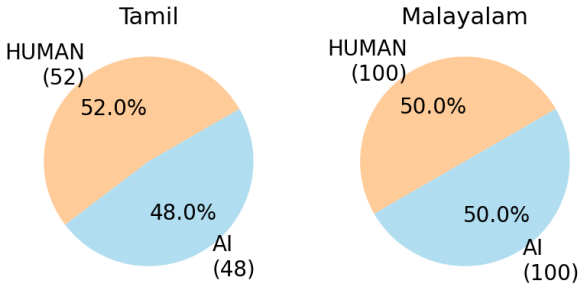


Figure 2: Distribution of labels in Tamil and Malayalam languages in given test dataset.

4 Methodology

4.1 Data Preprocessing

We did not require any data preprocessing steps for our dataset, as it consists of short texts with a max-

imum of 2 sentences and 3-4 words per sentence. Each of these sentences is clean and consistent, adhering to a standardized format. This high level of data quality means that there are no spelling errors, grammatical mistakes, or irrelevant content that would necessitate additional cleaning or normalization.

Furthermore, the language models we used to generate the embeddings, such as indic-bert (Kakwani et al., 2020), are designed to handle a variety of text inputs and perform certain preprocessing tasks internally. These models are capable of tokenizing the text, managing special characters and punctuation, and adjusting the length of text inputs through padding and truncation. This built-in preprocessing capability of the language models ensures that minor inconsistencies or noise in the data are effectively managed, allowing us to generate high-quality embeddings without the need for extensive data cleaning steps.

In summary, the combination of a clean and consistent dataset with the robust preprocessing capabilities of the language models we employed allowed us to bypass additional data preprocessing steps, streamlining our workflow and ensuring efficient and accurate text embedding generation.

4.2 Model Training

In our research, we aimed to detect AI-generated product reviews in Tamil and Malayalam by leveraging the strengths of monolingual models. We designed two monolingual models, one for Tamil and one for Malayalam.

First, we generated embeddings for each text entry in our dataset using a variety of language models, including XLM-RoBERTa (Conneau, 2019), Indic-BERT (Doddapaneni et al., 2023), and mT5 (Xue, 2020), which were trained on various languages, specifically on Tamil and Malayalam. Additionally, we employed Sentence-BERT (Reimers, 2019), which has been effectively used for AI-generated or AI-paraphrased text detection (Schaaff et al., 2024). These embeddings captured the semantic and syntactic properties of the text, providing a rich representation for further analysis.

To evaluate the effectiveness of each model's embeddings, we performed five-fold cross-validation using XGBoost. This ensured that our feature representations were robust across different subsets of the dataset.

Next, we split the dataset into three parts: 70% for training, 21% for testing, and 9% for validation. The training set was used to train the individual Deep Neural Network (DNN) models, while the testing set was used to evaluate their performance. We trained four separate DNN models independently using embeddings from each language model. To enhance overall performance, we employed a weighted average ensembling technique, leveraging the complementary strengths of different models.

Our evaluation metric was the F1-score, which provided a balanced measure of precision and recall, ensuring a more reliable assessment of classification performance compared to accuracy. By training the DNN with these features, we developed a streamlined and efficient model suitable for low-resource environments while maintaining strong classification performance in detecting AI-generated product reviews in Tamil and Malayalam.

Figure 3 illustrates the architecture of the model and figure 4 shows the detailed methodology of the work.

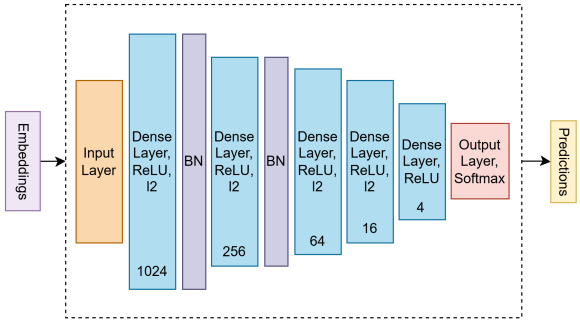


Figure 3: Proposed DNN architecture.

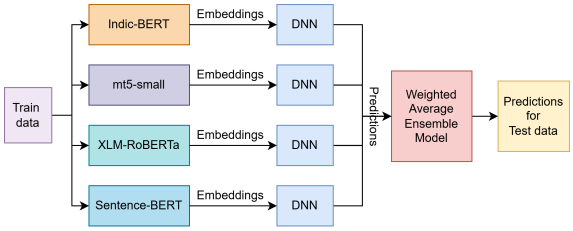


Figure 4: Proposed methodology.

5 Results and Discussion

Our initial XGBoost model achieved promising results on a 5-fold cross-validation set. Table 1 and Table 2 shows the mean and the standard deviation for each individual models. F1-score produced by

	Mean	Std.Dev
Sentence-BERT	0.962	0.014
XLM-RoBERTa	0.948	0.013
Indic-BERT	0.965	0.010
mT5	0.945	0.013

Table 1: Cross validation set mean and standard deviation for Tamil dataset.

	Mean	Std.Dev
Sentence-BERT	0.934	0.015
XLM-RoBERTa	0.909	0.010
Indic-BERT	0.927	0.011
mT5	0.930	0.010

Table 2: Cross validation set mean and standard deviation for Malayalam dataset.

the proposed DNN and ensemble models are shown in Table 3.

For the given test set, the ensemble model achieved 0.73 for Tamil and 0.90 for Malayalam. Respective confusion matrices are shown in the Figure 5 and Figure 6.

Despite achieving promising results with XGB on the cross-validation set and the proposed models on splitted test set, we observed a performance drop on the given test set for Tamil. This discrepancy may be attributed to differences in the distributions of the training and test sets, potentially generated by different LLM models. Future efforts will focus on refining the ensemble DNN model to ensure uniformity across varying distributions.

6 Conclusion

In this study, we explored AI-generated product review detection in Tamil and Malayalam using monolingual models with transfer learning and ensembling. Our approach achieved 90% accuracy for Malayalam and 73% for Tamil, demonstrating the effectiveness of transfer learning in low-resource Dravidian languages.

Models	Tamil	Malayalam
Sentence-BERT DNN	0.959	0.905
XLM-RoBERTa DNN	0.971	0.964
Indic-BERT DNN	0.971	0.935
mT5 DNN	0.953	0.964
Ensemble model	0.982	0.940

Table 3: Predictions of proposed models for 21% split of train dataset in Tamil and Malayalam.

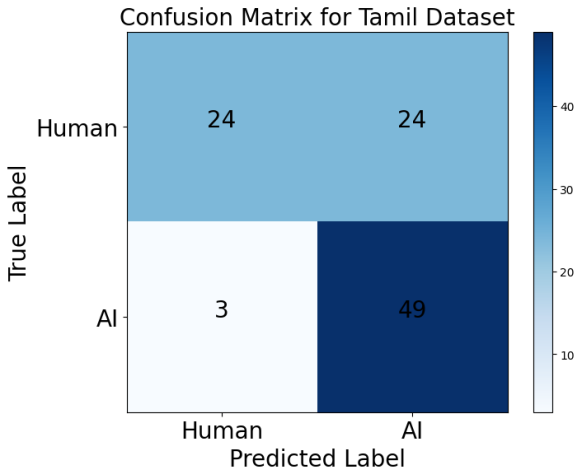


Figure 5: Confusion matrix for Tamil.

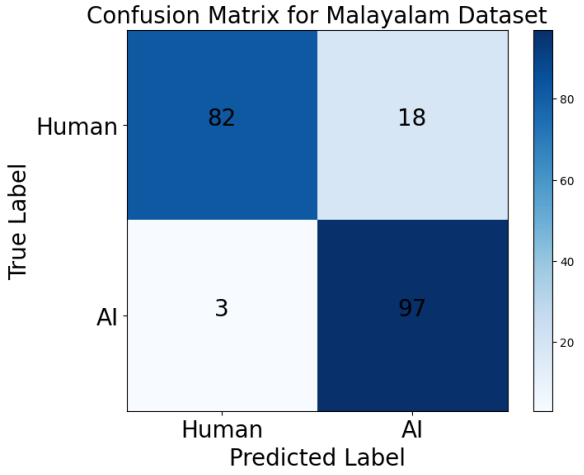


Figure 6: Confusion matrix for Malayalam.

To support research in this field, we have made our source code publicly available, enabling replication and further development. This contribution fosters innovation and collective efforts to enhance the reliability of AI-generated content detection, promoting the integrity of online reviews.

7 Limitations

The AI-generated reviews in the dataset may exhibit biases inherited from the language models used to generate them. These biases could affect the performance and fairness of the detection model, leading to variations in effectiveness. Additionally, the dataset is limited, which may further constrain the model's ability to learn. Addressing both dataset limitations and inherent biases remains a crucial area for future research.

References

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Zeynep Karaş. 2024. Effects of ai-generated misinformation and disinformation on the economy. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 12(4):2349–2360.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

George K Mikros, Athanasios Koursaris, Dimitrios Bilianos, and George Markopoulos. 2023. Ai-writing detection using an ensemble of transformers and stylometric features. In *IberLEF@ SEPLN*.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2024. Classification of human-and ai-generated texts for different languages and domains. *International Journal of Speech Technology*, 27(4):935–956.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

K Thilagavathi, K Thankamani, P Shunmugapriya, and D Prema. 2024. Navigating fake reviews in online marketing: Innovative strategies for authenticity and trust in the digital age. *The Scientific Temper*, 15(03):2854–2858.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2023. Gpt-who: An information density-based machine-generated text detector. *arXiv preprint arXiv:2310.06202*.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. Seqxgpt: Sentence-level ai-generated text detection. *arXiv preprint arXiv:2310.08903*.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural deepfake detection with factual structure of text. *arXiv preprint arXiv:2010.07475*.

DLRG@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages

Ratnavel Rajalakshmi, R. Ramesh Kannan, Meetesh Saini, Bitan Mallik
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, TamilNadu
rajalakshmi.r@vit.ac.in

Abstract

Social media is a powerful communication tool and rich in diverse content requiring innovative approaches to understand nuances of the languages. Addressing challenges like hate speech necessitates multimodal analysis that integrates textual, and other cues to capture its context and intent effectively. This paper proposes a multi-modal hate speech detection system in Tamil, which uses textual and audio features for classification. Our proposed system uses a fine-tuned Indic-BERT with Whisper as a multimodal approach for hate speech detection. The fine-tuned Indic-BERT model with Whisper achieved an F1 score of 0.25 on Multimodal based approach. Our proposed approach ranked at 10th position in the shared task on Multimodal Hate Speech Detection in Dravidian languages at the NAACL 2025 Workshop DravidianLangTech.

1 Introduction

Hate speech is a serious problem that harms people and communities. It targets individuals or groups based on characteristics like race, religion, or gender, leading to violence, discrimination, and prejudice. With the rise of social media, hate speech becomes more common, creating risks for society and individuals (Sreelakshmi et al., 2024; Wickramaarachchi et al., 2023; Khanduja et al., 2024). Many studies focus on detecting hate speech in widely spoken languages like English (Khanduja et al., 2024; Conneau et al., 2020), but Dravidian languages, such as Tamil, do not receive as much attention. Tamil is a complex language with a unique structure, making it important to develop specialized hate speech detection systems for it.

This paper presents a multimodal system to detect hate speech in Tamil. Hate speech appears in different forms, including text, audio, and images. Previous research explores transformer-based models like mBERT and XLM-R for hate speech detection (Khanduja et al., 2024; Ibañez et al., 2021). Some studies also develop hate speech detection systems for low-resource languages using multimodal approaches. To handle text and audio data, we propose a system that combines a fine-tuned Indic-BERT model with Whisper as a Multimodal hate speech classification. Our proposed method obtained a F1 score of 0.25 in detection of hate speech in Tamil.

2 Related Works

Hate speech detection has traditionally focused on monolingual text, with early approaches employing machine learning algorithms such as Support Vector Machines (SVMs) and n-gram features (Warner and Hirschberg, 2012). Such methods fail to deal with the subtle semantics of hate speech. Deep learning methods, such as CNNs and LSTMs, have achieved better results by encoding contextual information (Zhang et al., 2018). More recently, transformer models like BERT have further developed the area (Devlin et al., 2019). For other low resource languages like Tamil (Rajalakshmi et al., 2023; Ganganwar and Rajalakshmi, 2022), Marathi (Rajalakshmi et al., 2021a), Telugu (Rajalakshmi et al., 2024), Hindi (Rajalakshmi and Reddy, 2019; Rajalakshmi et al., 2021b) and Multilingual languages (Reddy and Rajalakshmi, 2020) authors have worked towards Hate and offensive content identification on textual data. Recent research has investigated multimodal methods for hate speech detection, especially utilizing both the text and audio modalities. (Mahajan et al., 2024) have used models that integrates CNNs and LSTMs to handle spectrograms and text embeddings for the identification of offensive speech. Likewise, (Khanduja et al., 2024) created dataset for hate speech detection in low resource Dravidian language Telugu. Explored transformer models such as mBERT, DistilBERT, IndicBERT, NLLB, Muril, RNN+LSTM, XLM-ROBERTa, and Indic-

Bart. Fine-tuned mBERT model achieved a accuracy of 98.2% on the newly created hate speech Telugu dataset. For sentiment analysis of tweets on social media content, (Kannan et al., 2021) employed the IndicBERT model. (Boishakhi et al., 2021) employed a multimodal approach for hate speech detection by combining video, audio, and transcribed text. For audio, features such as MFCC, ENERGY, ZCR, and chroma were utilized, while Bag of Words and TF/IDF were applied to transcribed text. A hard voting ensemble model was used to highlight the advantages of contextual analysis in achieving more accurate hate speech classification.

3 Model Architecture

3.1 Wav2Vec2 based model

Wav2Vec2 model is pre-trained on a large corpus of multilingual speech data. Model provides robust representations of acoustic features that benefits downstream tasks. The raw audio input, preprocessed is passed into the Wav2Vec2 model, which returns a sequence of contextualized representations of hidden states. Thus, the hidden states are able to capture subtle patterns within the audio and effectively encode temporal and spectral aspects of the signal. We then perform mean pooling over the time dimension on the sequence of hidden states obtained from Wav2Vec2. This pooling operation summarizes the audio input by aggregating all the temporal information into a fixed-length representation. The pooled representation is now given as a single vector that encapsulates the core acoustic features of the audio, and then it passes through a dropout layer. This layer introduces randomness while the model learns during training, reducing overfitting and improving the generalizability of the model to unseen data. The dropout rate, set at a specific value, regulates the number of neurons randomly deactivated during training. Finally, the output of the dropout layer feeds into a fully connected linear layer. This layer maps the pooled and regularized representation to a set of logits, corresponding to the five classes of our classification task. These logits during training are utilized to calculate cross-entropy loss when compared to the ground-truth labels allowing the model fine-tune on the task particular parameters. On inference, logits are the model output used as a way to determine the predicted class label using softmax activation functions to obtain a probability distribution of the classes.

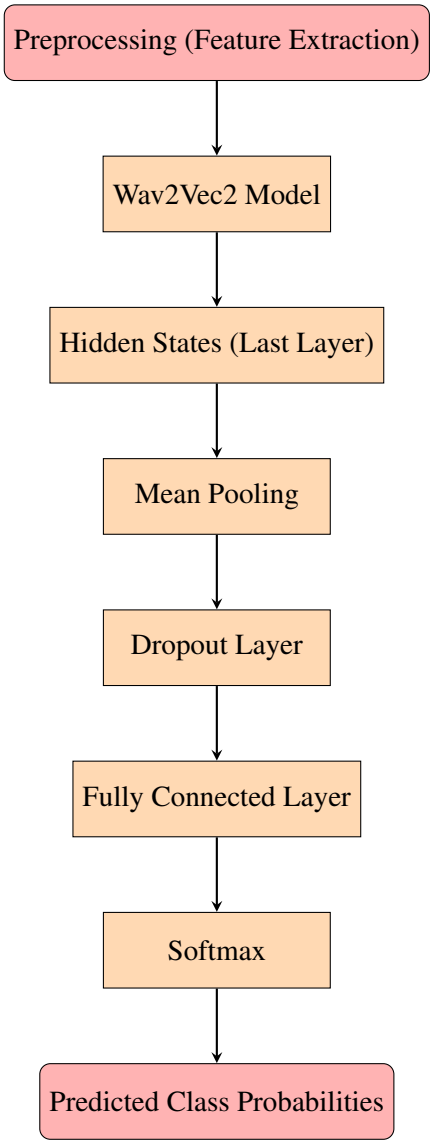


Figure 1: Flowchart of Wav2Vec2 Speech Classification Model

3.2 Indic-BERT+Whisper based model

Our proposed approach leverages Whisper for audio-to-text transcription, followed by IndicBERT for the classification of Telugu hate speech from the transcribed text. Whisper, a multilingual automatic speech recognition model, allows transcription of spoken Tamil so that subsequent textual analysis can be performed. Indic-BERT, a language model pre-trained on Indian languages exclusively, is superior to generic multilingual models like mBERT and XLM-R as it is able to capture linguistic variations specific to Tamil. This fusion overcomes the shortcomings of less specialized speech-to-text pipelines, making the transcriptions more appro-

priate for subsequent analysis. The core idea is the pooled output from the Indic-BERT model, which encapsulates the semantic understanding of the transcribed text. The pooled representation is passed through a dropout layer, which introduces stochasticity during training and prevents overfitting. The output of the dropout layer is then passed to a fully connected linear layer. This linear layer maps the contextualized text representation to the final output, which are the class logits. A cross-entropy loss is computed using the ground truth labels and predicted logits, which is used in model training. Finally, during inference, the logits are used to find class labels based on the highest probability. This architecture makes use of the capabilities of the Whisper model for speech-to-text conversion and Indic-BERT for contextual understanding in the classification task. Although Whisper is reported to have occasional transcription errors and it shows better performance in low-resource languages makes it a good choice. By integrating these two models, the proposed system efficiently handles both spoken and written Tamil, and it enhances the resilience of hate speech detection in a multimodal environment.

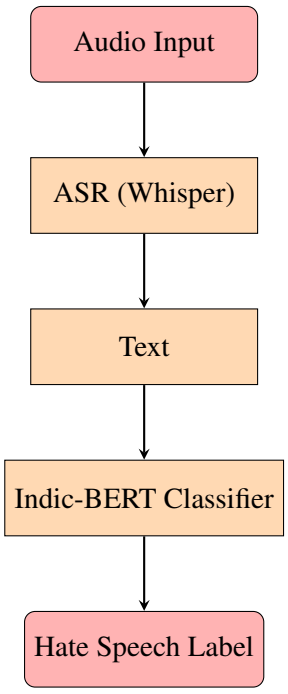


Figure 2: Flowchart of the Multimodal Hate Speech Detection System

4 Dataset Description

Multimodal Hate speech Telugu dataset contains text and audio data Tamil, Telugu and Malayalam. The audio samples are sourced from YouTube videos and encompass a variety of speakers, ensuring linguistic diversity and representation. Each utterance is labeled under one of five categories: N (Non-hate speech), G (Gender-based hate speech), P (Political hate speech), R (Religious hate speech), and C (Communal hate speech). Multimodal Hate speech Tamil dataset consists of 514 samples. It is well-structured into two categories, WAV files of speech recordings and transcriptions of text contents. The transcriptions are a textual representation of the spoken material, allowing for both linguistic analysis and machine learning use. The disbribution of the data is as shown in Figure 3. Additionally, a test dataset is provided, comprising 50 audio and text samples. However, the test labels are not included, requiring researchers to submit their model predictions for evaluation.

4.1 Data Preprocessing

The audio data is processed into raw audio signals, a format compatible with the feature extraction capabilities of the facebook/wav2vec2-large-xlsr-53 model. This process begins by loading raw audio files in .wav format and standardizing them by converting multi-channel audio to mono-channel, ensuring uniform input dimensionality. The preprocessing involves extracting the audio waveform and its corresponding sample rate. Since wav2vec2-large-xlsr-53 is trained on a 16 kHz sampling rate, input audio is resampled to match this rate. Instead of relying on manual feature engineering, the raw audio waveforms are directly passed through the model's feature extraction layers, enabling the model to learn relevant features autonomously. To meet the model's requirements, audio sequences are truncated to a maximum of 16,000 samples, equivalent to one second of audio sampled at 16 kHz, ensuring consistent input length. The audio processing pipeline in Indic-BERT employs a two-stage approach: transcription followed by classification. Initially, raw audio files are processed using the vasista22/whisper-tamil-medium model, configured specifically for Tamil transcription, to convert the audio into its textual representation with high accuracy. The resulting text undergoes preprocessing steps, including the removal of URLs and non-alphanumeric characters, as well as standardizing the input to a maximum length of 128 tokens through padding or truncation. The preprocessed text is then passed to the Indic-BERT model for classification.

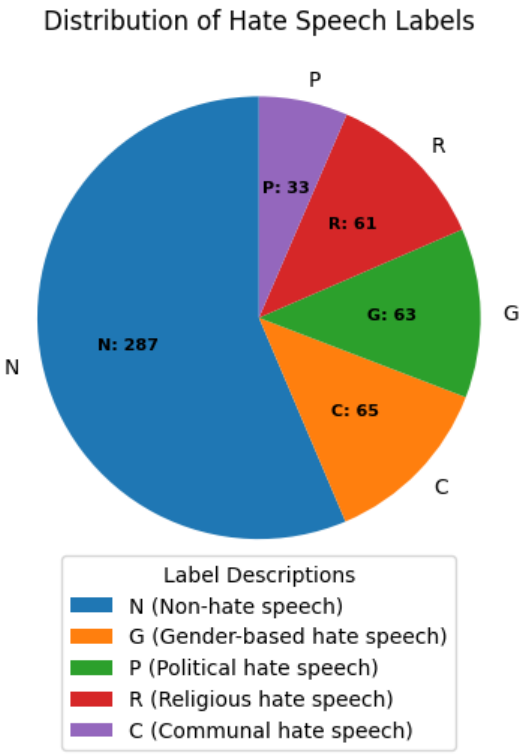


Figure 3: Distribution of Training dataset

5 Experiments and Results

The learning rate used to train both models is 3e-5 with the Adam optimizer. A linear warmup schedule is implemented during training, and both models are trained for a fixed number of epochs. At the end of each epoch, evaluation is performed to track the model progress. Gradient accumulation steps are used to optimize resource utilization for the Wav2Vec2 model. The model checkpoint with the highest held-out validation set accuracy is used in each case for the final model. For the Indic-BERT-based model, this is found at the 21st epoch, while for the Wav2Vec2-based model, training is conducted for 10 epochs. Wav2Vec2 and IndicBERT models were trained using a learning rate of 3e-5 with the Adam optimizer. The system is evaluated using a dataset from the shared task. The Wav2Vec2-based model attains an accuracy of 51% with an F1-score of 0.35 and Multimodal based Indic-BERT and Whisper obtained a F1 score of 0.25 on Tamil Hate Speech Detection. The following results were obtained after training the classification model:

Class	Prec	Recall	F1	Supp.
C	0.32	0.70	0.44	10
G	0.57	0.80	0.67	10
N	0.14	0.20	0.17	10
P	0.00	0.00	0.00	10
R	0.00	0.00	0.00	10

Metrics	Prec.	Recall	F1	Supp.
Accuracy			0.34	50
Macro Avg	0.21	0.34	0.25	50
Wei. Avg	0.21	0.34	0.25	50

Table 1: Classification report of Indic-BERT based model

6 Conclusion and Future Work

This paper explores multimodal hate speech detection in Tamil, using transcribed text and raw audio processed through separate architectures. A Whisper/Indic-BERT based multimodal approach captures textual and audio semantics and achieved a F1 score of 0.25. Wav2Vec2 focuses on only on speech features and obtained a F1 score of 0.35. Our findings laid the foundation for advanced models. The study highlights the importance of addressing challenges in audio-text integration and optimizing feature extraction. Future efforts will explore ensembling, dataset improvements, and enhanced audio pipelines, aiming to better integrate audio-text interactions for improved performance and societal impact.

References

Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md. Golam Rabiul Alam. 2021. Multi-modal hate speech detection using machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, page 4496–4499. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vaishali Ganganwar and Ratnavel Rajalakshmi. 2022. Mtdot: A multilingual translation-based data augmentation technique for offensive content identification in tamil text data. *Electronics*, 11(21):3574.

Michael Ibañez, Ranz Sapinit, Lloyd Antonie Reyes, Mohammed Hussien, Joseph Marvin Imperial, and Ramón Rodriguez. 2021. Audio-based hate speech classification from online short-form videos. In *2021 International Conference on Asian Language Processing (IALP)*, pages 72–77. IEEE.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

R Ramesh Kannan, Ratnavel Rajalakshmi, and Lokesh Kumar. 2021. Indicbert based approach for sentiment analysis on code-mixed tamil tweets. In *FIRE (Working Notes)*, pages 729–736.

Namit Khanduja, Nishant Kumar, and Arun Chauhan. 2024. Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation. *Systems and Soft Computing*, page 200112.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Esshaan Mahajan, Hemaank Mahajan, and Sanjay Kumar. 2024. Ensmulhatecyb: Multilingual hate speech and cyberbully detection in online social media. *Expert Systems with Applications*, 236:121228.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

R Rajalakshmi and B Yashwant Reddy. 2019. Dlrg@ hasoc 2019: An enhanced ensemble classifier for hate and offensive content identification.

Ratnavel Rajalakshmi, Faerie Mattins, S Srivarshan, Preethi Reddy, and M Anand Kumar. 2021a. Hate speech and offensive content identification in hindi and marathi language tweets using ensemble techniques. In *FIRE (Working Notes)*, pages 467–479.

Ratnavel Rajalakshmi, M Saptharishee, S Hareesh, R Gabriel, et al. 2024. Dlrg-dravidianlangtech@ eacl2024: Combating hate speech in telugu code-mixed text on social media. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 140–145.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Pavitra Vasudevan, et al. 2023. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech & Language*, 78:101464.

Ratnavel Rajalakshmi, S Srivarshan, Faerie Mattins, E Kaarthik, and Prithvi Seshadri. 2021b. Conversational hate-offensive detection in code-mixed hindi-english tweets. In *CEUR Workshop Proceedings*, pages 1–11.

Yashwanth Reddy and Ratnavel Rajalakshmi. 2020. Dlrg@ hasoc 2020: A hybrid approach for hate and offensive content identification in multilingual tweets. In *FIRE (working notes)*, pages 304–310.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

WAKM Wickramaarachchi, Sameeri Sathsara Subasinghe, KK Rashani Tharushika Wijerathna, A Sahashra Udani Athukorala, Lakmini Abeywardhana, and A Karunasena. 2023. Identifying false content and hate speech in sinhala youtube videos by analyzing the audio. In *2023 5th International Conference on Advancements in Computing (ICAC)*, pages 364–369. IEEE.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web*, pages 745–760, Cham. Springer International Publishing.

Team ML_Forge@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages

Adnan Faisal, Shiti Chowdhury, Sajib Bhattacharjee,
Udoy Das[†], Samia Rahman, Momtazul Arefin Labib, Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
[†]East Delta University, Bangladesh
{u2004002, u2004027, u2004003}@student.cuet.ac.bd, udoy.d@eastdelta.edu.bd,
{u1904022, u1904111}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Ensuring a safe and inclusive online environment requires effective hate speech detection on social media. While detection systems have significantly advanced for English, many regional languages, including Malayalam, Tamil and Telugu, remain underrepresented, creating challenges in identifying harmful content accurately. These languages present unique challenges due to their complex grammar, diverse dialects and frequent code-mixing with English. The rise of multimodal content, including text and audio, adds further complexity to detection tasks. The shared task "Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025" has aimed to address these challenges. A Youtube-sourced dataset has been provided, labeled into five categories: Gender (G), Political (P), Religious (R), Personal Defamation (C) and Non-Hate (NH). In our approach, we have used mBERT, T5 for text and Wav2Vec2 and Whisper for audio. T5 has performed poorly compared to mBERT, which has achieved the highest F1 scores on the test dataset. For audio, Wav2Vec2 has been chosen over Whisper because it processes raw audio effectively using self-supervised learning. In the hate speech detection task, we have achieved a macro F1 score of 0.2005 for Malayalam, ranking 15th in this task, 0.1356 for Tamil and 0.1465 for Telugu, with both ranking 16th in this task.

1 Introduction

With the increasing spread of harmful content online, hate speech detection on social media has become a crucial area of research. While platforms empower users to express views, they are often exploited to propagate hate and abuse. Despite progress in English, regional languages like Malayalam, Tamil and Telugu remain under-researched, highlighting the need for more inclusive detection frameworks. These Dravidian languages, spoken in southern India, present challenges such as complex grammar, dialect variations and code-mixing (Chakravarthi et al., 2022). Existing research has faced gaps due to the lack of large, well-balanced datasets, limiting robust machine learning models (Premjith et al., 2024a). Most studies have focused on text analysis, ignoring the multimodal nature of social media content, which includes both text and audio (Chakravarthi et al., 2021).

The shared task "Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025" has addressed key challenges in this area (Lal G et al., 2025). The dataset, sourced from YouTube, requires models to analyze both text and audio components for detecting hate speech in Malayalam, Tamil and Telugu.

In this study, we have proposed a multimodal approach using mBERT for text and Wav2Vec2 for audio. The hybrid mBERT + Wav2Vec2 model has shown improved performance, achieving F1 scores of 0.3013 for Malayalam, 0.2853 for Tamil and 0.2511 for Telugu, demonstrating the effectiveness of combining textual and acoustic features (Premjith et al., 2024b). This approach has emphasized the benefits of multimodal fusion in overcoming the limitations of single-modality models and advancing hate speech detection in underrepresented languages (B et al., 2024). The core contributions of our research work are as follows -

- We have used augmentation Technique to balance dataset for Malayalam, Tamil and Telugu languages.
- We have utilized an efficient fusion technique to improve classification and enhance model performance.

Detailed implementation information is available in the GitHub repository - https://github.com/Sojib001/MHDS

2 Related Work

The rapid growth of social media has led to increased hate speech and abusive content, raising concerns about platform safety. While progress has been made in hate speech detection for high-resource languages like English, under-resourced languages such as Malayalam, Tamil and Telugu face challenges due to complex grammar, dialects, frequent English mixing and limited labeled data (Chakravarthi et al., 2021).

Unimodal approaches, relying on text-based models like BERT (Devlin et al., 2019; Liu et al., 2019) and mBERT, have analyzed linguistic features. Sreelakshmi et al. (2024) has addressed dataset imbalances in Dravidian languages and Chakravarthi et al. (2021) has demonstrated the effectiveness of transformer models for Tamil and Malayalam. However, unimodal methods have lacked the ability to incorporate signals from other modalities.

Multimodal Approaches: Combining text and audio has shown promise in hate speech detection by capturing linguistic and acoustic cues. B et al. (2024), Kiela et al. (2020) and (Anilkumar et al., 2024) have demonstrated improved detection accuracy using multimodal approaches. Chakravarthi et al. (2021) has highlighted the potential of YouTube-sourced multimodal datasets. Despite this progress, research on multimodal hate speech detection for Dravidian languages remains limited, requiring further exploration.

3 Data Description

The dataset for Multimodal Hate Speech Detection in Malayalam, Tamil and Telugu is sourced mainly from YouTube. It includes text from captions and audio from videos, covering both speech and background noise. Hate speech is categorized into Gender (G), Political (P), Religious (R), Personal Defamation (C) and Non-Hate (NH). Table 1 presents the dataset distribution across training and test sets.

Language	Training Dataset	Test Dataset
Malayalam	883	50
Tamil	514	50
Telugu	556	50
Total	**1953**	**150**

Table 1: Language-wise Distribution of Training and Test Data

Table 2 shows the distribution of five class labels—Gender (G), Political (P), Religious (R), Personal Defamation (C) and Non-Hate (NH)—across Malayalam, Tamil and Telugu, with total counts at the bottom.

Class Label	Malayalam	Tamil	Telugu
Gender (G)	82	101	63
Political (P)	118	58	33
Religious (R)	91	72	61
Personal Defamation (C)	186	122	65
Non-Hate (NH)	406	198	287
Total	**883**	**514**	**556**

Table 2: Statistical Breakdown of Class Labels Across Malayalam, Tamil and Telugu

4 Methodology

4.1 Problem Formulation

The task has been to detect hate speech in Malayalam, Tamil and Telugu across five categories: Gender, Political, Religious, Personal Defamation and Non-Hate. We have used late fusion to combine text features from mBERT and audio features from Wav2Vec2, merging them in a classification layer to improve predictions, despite challenges like language differences and code-mixing.

4.2 Data Augmentation and Preprocessing

To address class imbalance, data augmentation was applied to balance the dataset. Back translation expanded the training data by translating text to another language and back, creating variations. For audio, irrelevant features like MFCC were removed to clean the data and focus on useful information.

4.3 Uni-modal Models

4.3.1 Text-based Model

We used mBERT (multilingual BERT) and T5 for text classification due to their ability to capture contextual meanings across languages. mBERT, fine-tuned on Malayalam, Tamil and Telugu datasets, outperformed T5 in effectively classifying hate speech.

4.3.2 Audio-based Model

We have used Wav2Vec2 and Whisper for audio feature extraction, with Wav2Vec2 performing better in capturing tone, pitch and context for hate speech detection. Initially, MFCC features were included but have been removed after receiving the gold test data, improving performance with only Wav2Vec2 features.

4.4 Fusion Model

To enhance classification accuracy, we have adopted a multimodal fusion approach that integrates textual and audio features through late fusion, where mBERT (text) and Wav2Vec2 (audio) representations are combined for improved hate speech detection. mBERT has consistently outperformed T5 in text processing, while Wav2Vec2 has excelled over Whisper in capturing audio features. By merging mBERT's text embeddings and Wav2Vec2's audio features, our model effectively captures both linguistic and acoustic nuances, leading to a more robust and accurate detection system. Figure 1 illustrates the overall modeling pipeline, showcasing the integration of text and audio features through the late fusion mechanism.

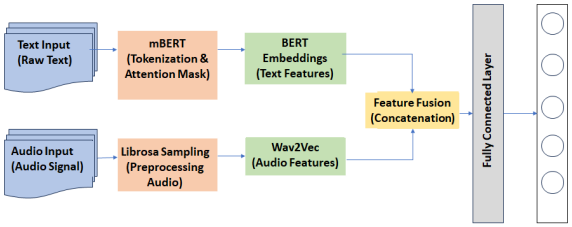


Figure 1: Abstract process of violence text detection

4.5 Evaluation Metrics

The models were evaluated using macro-F1 score, precision and recall to ensure balanced performance and accurate identification of hate speech.

5 Results and Analysis

This task has evaluated models for detecting hate speech in Malayalam, Tamil and Telugu using both text and audio data. The results have shown good performance on training data but struggles with test data, highlighting issues like overfitting, class imbalance and challenges in combining text and audio data.

5.1 Task 1: Malayalam Multimodal Hate Speech Detection

Table 3 has shown the performance of different classifiers for Malayalam. Among text-based models, mBERT has achieved the highest F1 score (0.5796), outperforming T5 (0.45). For audio models, Wav2Vec2 has performed best (F1: 0.3399), surpassing Whisper (0.30). In multimodal setups, mBERT + Wav2Vec2 have achieved the highest F1 score (0.3013), demonstrating the effectiveness of combining text and audio features. Figure 2a has represented the confusion matrix of our best-performing model.

Malayalam	Classifier	P	R	F1
Unimodal (Text)	mBERT	0.64	0.61	0.57
	T5	0.42	0.42	0.45
Unimodal (Audio)	Wav2Vec2	0.31	0.34	0.33
	Whisper	0.27	0.34	0.30
Multi-modal	**(mBERT + Wav2Vec2)**	**0.31**	**0.30**	**0.30**
	(T5 + Wav2Vec2)	0.21	0.28	0.24
	(mBERT + Whisper)	0.30	0.29	0.26

Table 3: Performance of Malayalam Classifiers (Macro Average)

5.2 Task 2: Tamil Multimodal Hate Speech Detection

Table 4 shows the classification performance for Tamil, where mBERT has achieved the highest F1 score (0.5561) for text. In the audio category, Whisper has reached an F1 score of 0.1494. The multimodal setup, combining mBERT with Wav2Vec2, has achieved an F1 score of 0.2853, demonstrating the benefits of integrating text and audio features. This fusion model has effectively combined mBERT's text embeddings with Wav2Vec2's speech representations. Figure 2b shows the confusion matrix of the best-performing model.

Tamil	Classifier	P	R	F1
Unimodal (Text)	mBERT	0.62	0.56	0.55
	T5	0.48	0.52	0.42
Unimodal (Audio)	Wav2Vec2	0.13	0.16	0.14
	Whisper	0.13	0.17	0.14
Multi-modal	**(mBERT + Wav2Vec2)**	**0.34**	**0.30**	**0.29**
	(T5 + Wav2Vec2)	0.32	0.28	0.27
	(mBERT + Whisper)	0.32	0.29	0.28

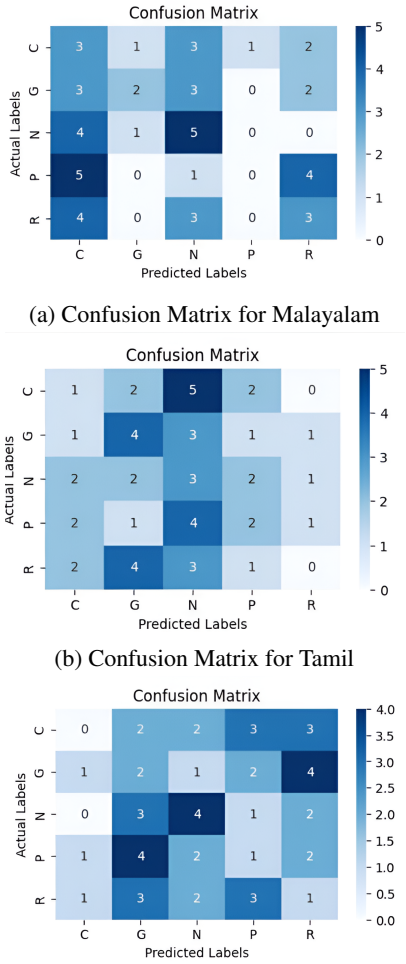
Table 4: Performance of Tamil Classifiers (Macro Average)

5.3 Task 3: Telugu Multimodal Hate Speech Detection

As shown in Table 5, different models perform differently for Telugu. mBERT achieves the best F1 score among text models at 0.3176. For audio models, Wav2Vec2 and Whisper score 0.1790 and 0.1894, respectively. Among combined models, (mBERT + Wav2Vec2) performs the best with an F1 score of 0.2511. Figure 2c represents the confusion matrix of our best performing model that combines mBERT with Wav2Vec2.

Telugu	Classifier	P	R	F1
Unimodal (Text)	mBERT	0.32	0.34	0.31
	T5	0.31	0.33	0.29
Unimodal (Audio)	Wav2Vec2	0.16	0.20	0.17
	Whisper	0.17	0.14	0.18
Multi-modal	**(mBERT + Wav2Vec2)**	**0.25**	**0.26**	**0.25**
	(T5 + Wav2Vec2)	0.21	0.25	0.19
	(mBERT + Whisper)	0.21	0.24	0.23

Table 5: Performance of Telugu Classifiers (Macro Average)



(a) Confusion Matrix for Malayalam

(b) Confusion Matrix for Tamil

(c) Confusion Matrix for Telugu

Figure 2: Confusion Matrices for Malayalam, Tamil and Telugu

We have selected mBERT and Wav2Vec2 for their strong text and audio capabilities. mBERT, pretrained on Malayalam, Tamil and Telugu, has outperformed T5, while Wav2Vec2 has excelled over Whisper. Their integration has improved classification accuracy. To address overfitting, we have applied early stopping (patience = 5), L1 regularization and hyperparameter tuning. However, back translation has degraded performance.

The confusion matrices have shown classification trends, with off-diagonal elements revealing misclassifications. Figure 2 presents error patterns like C $\leftrightarrow$ N, P $\leftrightarrow$ R and G $\leftrightarrow$ N. Malayalam has confused P and R, Tamil C and N and Telugu G and N, highlighting challenges in distinguishing sentiment and contextual labels.

5.4 Parameter Setting

Our best-performing model, mBERT, has used a learning rate of 0.00002, batch size 8 and the Adam optimizer, with a text input size of 512 and early stopping (patience = 5) to prevent overfitting. Wav2Vec2 has applied the same learning rate and batch size, while Whisper used a learning rate of 0.00001. T5 has followed mBERT's settings but performed worse. As shown in Table 6, the Fusion Model has combined mBERT's text features with Wav2Vec2's audio, improving multimodal classification in Malayalam, Tamil and Telugu. .

Model	Learning Rate	Optimizer	Batch Size
mBERT	2e-5	AdamW	8
Wav2Vec2	2e-5	AdamW	8
Whisper	1e-5	AdamW	8
T5	2e-5	AdamW	8
Fusion Model	2e-5	Adam	8

Table 6: Key Hyperparameters for Model Training

6 Conclusion

The Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025 has identified key challenges in detecting hate speech in Malayalam,

Tamil and Telugu using text and audio. While transformer models have been effective, they have faced overfitting, data imbalance and multimodal integration issues. mBERT and Wav2Vec2 have shown overfitting, excelling in training but underperforming in testing. Multimodal fusion has shown promise but has struggled with noisy audio, alignment issues and class imbalance. Despite these challenges, the fusion model has been effective, leveraging mBERT's text embeddings and Wav2Vec2's speech representations to enhance classification. However, synchronization issues and noisy audio have impacted performance. Regarding back translation, it has not affected code-mixing patterns but has occasionally produced unnatural sentences, requiring manual validation. Since code-mixed texts have remained intact, their linguistic integrity has been preserved.

Limitations

The main limitation of our model is it has overfitted. It has learned noise and specific patterns in the training set that don't generalize. Not having enough training data also led our model to poor generalization. It performs worse than the unimodals in this task.

Ethical Statement

All data processing and modeling followed ethical rules for dealing with sensitive information, such as hate speech. The study seeks to improve hate speech detection while protecting rights and privacy of the people. The goal of the results is to improve moderation on online platforms and create safer spaces for users. We have recognized and handled any biases or limitations in the dataset as much as we could.

Acknowledgement

This document has been adapted from prior ACL and NAACL proceedings, including NAACL 2025 guidelines by Lal G. Jyothish, Premjith B. and Bharathi Raja Chakravarthi. We appreciate contributions from NAACL 2024 and ACL 2023 by Premjith B., Chakravarthi Bharathi Raja, Saranya Rajiakodi, Shubhankar Barman and Mithun Das. We express gratitude to the DravidianLangTech workshop organizers for promoting research in Dravidian languages. We also value insights from IEEE Access 2024 by Sreelakshmi K., Premjith B. and Bharathi Raja Chakravarthi, along with studies by Anilkumar et al. (2024), Mithun Das and Kiela et al. (2020) on multimodal abusive language detection. We acknowledge contributions from Chakravarthi et al. (2021, 2022) and Thapa et al. (2022) in multimodal sentiment analysis. Special appreciation to ACL and the NLP community for their dedication to dataset development and research on under-resourced languages.

References

Abhishek Anilkumar, Jyothish Lal G, B Premjith, and Bharathi Raja Chakravarthi. 2024. Dravlangguard: A multimodal approach for hate speech detection in dravidian social media. In *Speech and Language Technologies for Low-Resource Languages (SPELLL)*, Communications in Computer and Information Science.

Premjith B, Jyothish G, Sowmya V, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, Bharathi B, Saranya Rajiakodi, Rahul Ponnusamy, Jayanth Mohan, and Mekapati Reddy. 2024. Findings of the shared task on multimodal social media data analysis in Dravidian languages (MSMDA-DL)@DravidianLangTech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Jishnu Parameswaran P. K, Premjith B, K. P Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, and John P. McCrae. 2021. Dravidianmultimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *Preprint*, arXiv:2106.04853.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2022. Dravidiancodemix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview

of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

codecrackers@DravidianLangTech 2025: Sentiment Classification in Tamil and Tulu Code-Mixed Social Media Text Using Machine Learning

Lalith Kishore V P [1], Manikandan G [1], Mohan Raj M A[1]
Keerthi Vasan A[1], Aravindh M [1]
[1]R.M.K. Engineering College, Tiruvallur, Tamilnadu, India
{lali22025, gmk, moha22029, keer22061, arav22001}.ad@rmkec.ac.in

Abstract

Sentiment analysis of code-mixed Dravidian languages has become a major area of concern with increasing volumes of multilingual and code-mixed information across social media. This paper presents the "Seventh Shared Task on Sentiment Analysis in Code-mixed Tamil and Tulu", which was held as part of DravidianLangTech(NAACL-2025). However, sentiment analysis for code-mixed Dravidian languages has received little attention due to issues with class imbalance, small sample size, and the very informal nature of the code-mixed text. This study applied an SVM-based approach for the sentiment classification of both Tamil and Tulu languages. The SVM model achieved competitive macro-average F1 scores of 0.54 for Tulu and 0.438 for Tamil, showing that traditional machine learning methods can address the problem of sentiment categorization in code-mixed languages under low-resource settings.

1 Introduction

Sentiment analysis consists of classifying text depending on the opinions and emotions of the writer expressed inside it. The DravidianLangTech shared task is meant to further research in this field by concentrating on code-mixed datasets of comments and posts, especially in low-resource languages such as Tamil and Tulu (S. K. et al., 2024). This task is a need of the day when social media platforms such as Instagram, X and YouTube serve as the principal portals for communication across the world, being used by a mass of people to express their opinions and emotions in a cross-linguistic or cross-geographical manner. This shared task helps develop robust sentiment classification models and establishes a benchmark for evaluating techniques in linguistically diverse settings. It addresses challenges in informal and code-mixed language use, contributing to more inclusive and effective sentiment analysis for low-resource languages. The rest of the paper is organized as follows. Section 2 outlines the related works emphasizing Sentiment Analysis in Dravidian languages. Section 3 presents a description of the dataset and data processing. Section 4 describes the methodology used for the shared task. Section 5 discusses the result and findings of the task assigned. In Section 6 concludes the paper. Section 7 highlights the limitations of this study. At last, we have the references.

2 Related Work

Sentiment analysis (SA) in code-mixed Dravidian languages has gained momentum with the rise of multilingual social media content. Foundational datasets like (Chakravarthi et al., 2020) for Tamil and (Hegde et al., 2022) for Tulu have enabled systematic exploration of code-mixed SA, though challenges persist in class imbalance, informal text, and low-resource settings. Several ML models are experimented with various features for SA of user-generated content in code-mixed low-resource languages (Hegde et al., 2023). Initial studies used standard machine learning (ML) models and applied feature extraction techniques like TF-IDF and Bag-of-Words (BoW). (Shanmugavadivel et al., 2024a) investigated Decision Trees, along with SVM, using Tamil code-mixed data; results showed high accuracy (99%) but low macro F1-scores (0.39), indicating a class imbalance problem. Like (B et al., 2024) a large ensemble of machine learning models with careful optimization was used, achieving considerably better macro F1-scores of 0.26 (Tamil) and 0.55 (Tulu). However, a closer look at the confusion matrices showed difficulties in distinguishing between subtle sentiments like "Mixed Feelings" and others. Several conventional machine learning approaches exhibit mean-

https://github.com/VPLALITHKISHORE/DravidianLangTech_SharedTask

ingful limitations when confronted with the informal structure and skewed data distribution characteristic of code-mixed text, as these studies clearly show. Bi-LSTM and transformer architectures address contextual nuances. Roy and Kumar (2021) combined GloVe embeddings with Bi-LSTM for Tamil, achieving a weighted F1-score of 0.552 but faltering on minority classes. (Tripty et al., 2024) leveraged XLM-RoBERTa for Tulu (F1: 0.468) and used back-translation for Tamil, underscoring transformers potential despite data limitations. Class imbalance remains critical. (Kanta, 2023) observed F1-scores as low as 0.147 for Tamil using SVM, while (Shanmugavadivel et al., 2024a) noted Decision Trees 99% accuracy but 0.39 F1 due to skewed distributions. Recent solutions include data augmentation (Tripty et al., 2024) and hard-voting ensembles (B et al., 2024). (Ponnusamy et al., 2023) proposed ML models (LR, Multinomial Naive Bayes (MNB), and LinearSVC) trained with Term Frequency-Inverse Document Frequency (TF-IDF) of word unigrams for SA in Tamil and Tulu languages. Their proposed LR, MNB, and LinearSVC models obtained macro F1 scores of 0.43, 0.20, 0.41 and 0.51, 0.25, 0.49 for Tamil and Tulu languages respectively.

3 Dataset resource and data processing

To analyze sentiment in code-mixed languages, we leveraged existing datasets curated for sentiment analysis such as (Chakravarthi et al., 2020) for Tamil and (Hegde et al., 2022) for Tulu.

Labels	Train Set	Development Set	Test Set
Positive	18145	2272	1983
unknown_state	5164	619	593
Negative	4151	480	458
Mixed_feelings	3662	472	425
Total	31122	3843	3459

Table 1: Label-wise Breakdown of Tamil Code-Mixed Data.

The preprocessing phase involved standardizing and cleaning text data for both Tulu and Tamil languages to ensure consistency and reduce noise. While maintaining Tulu script characters using their Unicode range, text normalisation for Tulu involved changing all characters to lowercase and removing non-alphanumeric symbols. Similar normalisation was applied to Tamil text, with particular focus on keeping the characters from the Tamil script. Stopwords were eliminated by combining lists of English and language specific stopwords. Tamil used a comprehensive predefined list of stopwords, whereas Tulu used a custom-curated list. Using TF-IDF vectorization, feature extraction was carried out. Tamil employed unigrams just for simplicity, whereas Tulu used bigrams and unigrams to capture contextual subtleties. To ensure linguistic integrity, script-specific characters were kept intact throughout the tokenization process.

Labels	Train Set	Development Set	Test Set
Not Tulu	4400	543	474
Positive	3769	470	453
Neutral	3175	368	343
Mixed	1114	143	120
Negative	843	118	88
Total	13301	1642	1478

Table 2: Label-wise Breakdown of Tulu Code-Mixed Data.

4 Methodology

The methodology applied supports vector machines (SVMs), which were selected for this specific application because of their reliability with high-dimensional text data. In Tulu, class imbalance was countered with the generation of synthetic examples of the minority class via SMOTE oversampling, while hyper-parameter tuning through grid-search was applied to find optimal kernel and regularization parameters for the model. SMOTE was chosen over ADASYN because it generates synthetic samples evenly across the minority class, ensuring balanced augmentation without amplifying noise. ADASYN, which focuses on harder-to-learn examples, can introduce unwanted noise and overfitting, especially with high-dimensional TF-IDF features. SMOTE provides better stability for text classification. Conversely, the Tamil model used Class Weight Adjustment. Both models transformed text to TF-IDF vectors, Tulu's vectorizer being more oriented toward n-gram associations. The validation measures used included accuracy, precision, recall, and F1-scores-with Tulu's macro-averaged AUC-ROC being applied to assess multi-class performance. Classification reports contained performance details, sentiment distribution maps illustrated class balances, and confusion matrices revealed the patterns behind misclassifications; hence reproducibility and deployment readiness were pro-

vided by applying the optimized models in making the final predictions upon the test data with results stored for both languages.

4.1 Feature Extraction

The Tulu and Tamil datasets feature extraction procedure focused on the most informative words for sentiment classification by converting raw text into numerical vectors using the TF-IDF method. In order to capture both individual words and contextual phrases that may be essential for sentiment expression, such as negations or emotive combinations, both unigrams and bigrams (ngram_range=(1, 2)) were employed for the Tulu dataset. The feature space was limited to the top 5,000 most frequent terms (max_features=5000), allowing for efficient computation while retaining the most significant features. Furthermore, rare terms (those that appeared in fewer than three documents) were filtered out for the Tulu dataset using min_df=3, helping to eliminate noise.

4.2 Models and Techniques Utilized

For the Tulu task, an SVM model with GridSearchCV-optimized hyperparameters (kernel type, regularization C) was implemented. Class imbalance was addressed via SMOTE (synthetic oversampling). Features were extracted using TF-IDF (unigrams/bigrams, min_df=3) to filter rare terms. Performance was evaluated via macro-F1 (handling imbalance) and AUC-ROC (multi-class). For the Tamil task, a Linear SVM (C=1) with class weight adjustment was used to handle class imbalance instead of oversampling. TF-IDF focused on codemixed tokens (Tamil-English keywords). Metrics included accuracy and weighted F1 to assess sentiment across imbalanced classes. This can be observed with the help of visuals as in Table 3.

Component	Tulu	Tamil
Feature Extraction	Unigrams+bigrams, min-df=3	Unigrams only, codemix-aware tokens
Model	GridSearch-optimized	Linear SVM(fixed C=1)
Class Balancing	SMOTE oversampling	Class Weight Adjustment

Table 3: Label-wise Breakdown of Tulu and Tamil Code-Mixed Data

4.3 Classification for Tamil and Tulu codemix

The ability of the Tamil model to differentiate across various labels is demonstrated by the matrix. This can be observed with the help of visuals as in Figure 1. The ability of the Tulu model to differentiate across various labels is demonstrated by the matrix. This can be observed with the help of visuals as in Figure 2.

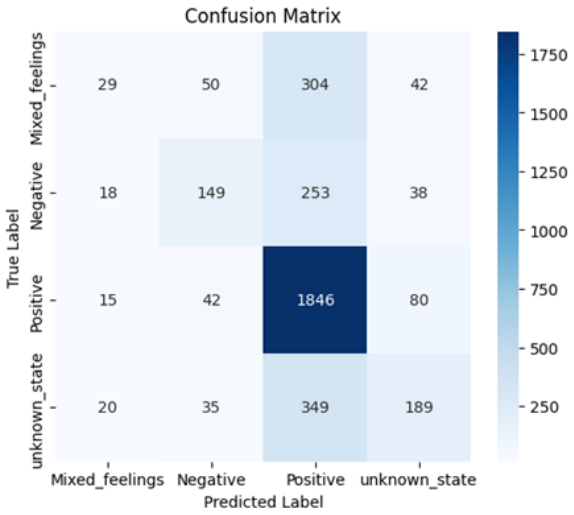


Figure 1: Confusion matrix of the proposed model for code-mixed Tamil text

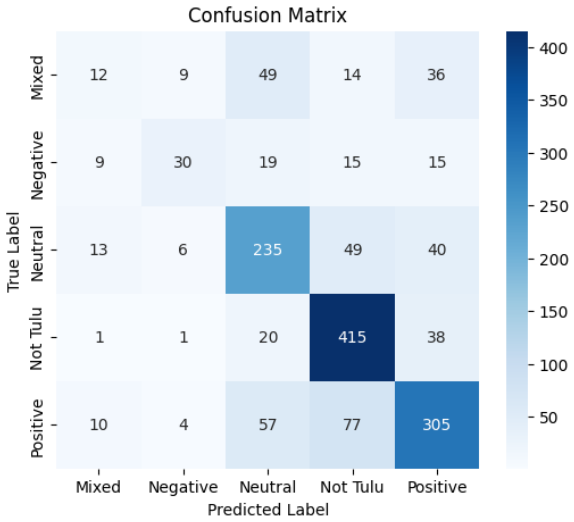


Figure 2: Confusion matrix of the proposed model for code-mixed Tulu text

5 Result and Findings

The performance of the model was evaluated across the labels based on the task. The models outperformed with this approach, achieving competitive macro average F1-scores of **0.54** for Tulu and **0.438** for Tamil. The sentiment distribution comparison of the labels and the classification report for Tamil can be observed with the help of visuals in Figure 3.

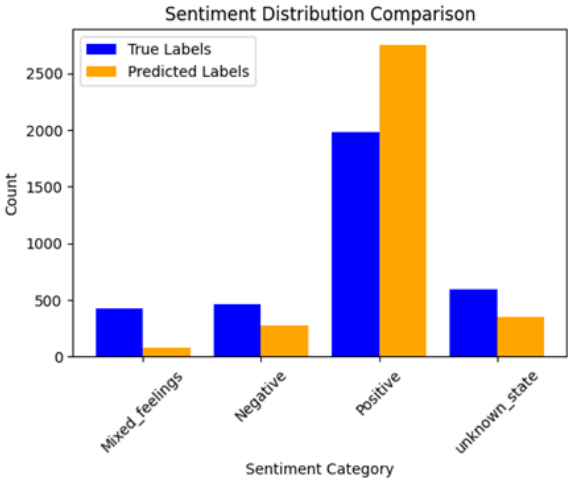


Figure 3: Sentiment Distribution Comparison and Classification report table for code-mixed Tamil.

The sentiment distribution comparison of the labels and the classification report for Tulu can be observed with the help of visuals in Figure 4.

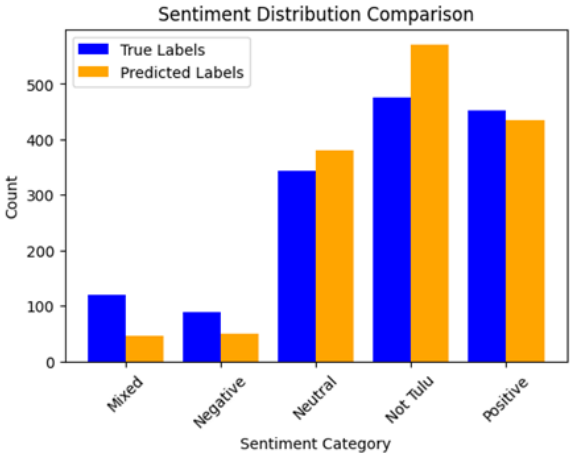


Figure 4: Sentiment Distribution Comparison and Classification report table for code-mixed Tulu.

6 Conclusion

SVM has proven to be a robust and adaptable approach for NLP tasks, particularly in low-resource settings. By leveraging hyperparameter tuning, SMOTE, and TF-IDF with n-grams, the model effectively handled class imbalance and noise in the Tulu dataset, achieving strong macro-F1 and AUC-ROC scores. For Tamil codemixed data, a Linear SVM with class weight adjustments and TF-IDF-based feature selection provided stable performance, as reflected in the weighted F1 score and accuracy. These results highlight the importance of tailoring preprocessing and optimization strategies to dataset characteristics—extensive balancing techniques for highly skewed distributions, and minimal interventions for well-represented codemixed contexts. Overall, SVM remains a powerful choice for sentiment classification, demonstrating its effectiveness in diverse linguistic and data imbalance scenarios.

7 Limitations

Tamil and Tulu SVM-based sentiment classification models face several limitations. The Tamil model, with a fixed linear SVM and class weight adjustment, may suffer from underfitting and struggles with colloquial expressions, transliterated words (Tanglish) and phrase-level sentiment detection due to its unigram-based TF-IDF approach. Although it partially handles Tanglish, it fails to capture complex code-mixed structures, sentiment shifts, and negations effectively. The Tulu model, while using SMOTE for imbalance correction, may introduce synthetic noise and relies on a manual stopword list, which might not fully cover dialectal variations.

References

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Asha Hegde, Mudoor Devadas Anusha, Sharal

Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Lavanya S K, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Durairaj Thenmozhi, and Rajkumar Charmathi Kumaresan, Prasanna Kumar. 2024. Overview of Second Shared Task on Sentiment Analysis in Code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, SK Lavanya, Durairaj Thenmozhi, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64-71.

Kogilavani Shanmugavadivel, Sowbharanika Janani J S, Navbila K, and Malliga Subramanian. 2024a. Code Maker@DravidianLangTech-EACL 2024: Sentiment Analysis in Code-Mixed Tamil using Machine Learning Techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Prathvi B, Manavi K K, Subrahmanya, Asha Hegde, Kavya G, and H L Shashirekha. 2024. MUCS@DravidianLangTech-2024: A Grid Search Approach to Explore Sentiment Analysis in Codemixed Tamil and Tulu . In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Pradeep Kumar Roy, and Abhinav Kumar. "Sentiment Analysis on Tamil Code-Mixed Text using Bi-LSTM." FIRE (Working Notes). 2021.

Selam Abitte Kanta and Grigori Sidorov. 2023. Selam@DravidianLangTech:Sentiment Analysis of Code-Mixed Dravidian Texts using SVM Classification.In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria.

Zannatul Fardaush Tripty, Md. Arian Al Nafis, Antu Chowdhury, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUETSentimentSillies@DravidianLangTechEACL2024: Transformer-based Approach for Sentiment Analysis in Tamil and Tulu Code-Mixed Texts. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Kishore Kumar Ponnusamy, Charmathi Rajkumar, Prasanna Kumar Kumaresan, Elizabeth Sherly, and Ruba Priyadharshini. 2023. VEL@ DravidianLangTech: Sentiment Analysis of Tamil and Tulu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 211–216. Varna, Bulgaria. Recent Advances in Natural Language Processing.

CUET_Ignite@DravidianLangTech 2025: Detection of Abusive Comments in Tamil Text Using Transformer Models

MD. Mahadi Rahman , Mohammad Minhaj Uddin and **Mohammad Shamsul Arefin**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904094, u1904118}@student.cuet.ac.bd, sarefin@cuet.ac.bd

Abstract

Abusive comment detection in low-resource languages is a challenging task particularly when addressing gender-based abuse. Identifying abusive language targeting women is crucial for effective content moderation and fostering safer online spaces. A shared task on abusive comment detection in Tamil text organized by DravidianLangTech@NAACL 2025 allowed us to address this challenge using a curated dataset. For this task, we experimented with various machine learning (ML) and deep learning (DL) models including Logistic Regression, Random Forest, SVM, CNN, LSTM, BiLSTM and transformer-based models such as mBERT, IndicBERT, XLM-RoBERTa and many more. The dataset comprised of Tamil YouTube comments annotated with binary labels, Abusive and Non-Abusive capturing explicit abuse, implicit biases and stereotypes. Our experiments demonstrated that XLM-RoBERTa achieved the highest macro F1-score(0.80), highlighting its effectiveness in handling Tamil text. This research contributes to advancing abusive language detection and natural language processing in low-resource languages particularly for addressing gender-based abuse online.

1 Introduction

Social media platforms have become integral to modern communication, offering spaces for individuals to express opinions, share experiences and engage in public discourse. However these platforms are also increasingly plagued by abusive content including hate speech, harassment and gender-based violence (Pannerselvam et al., 2023). Among the many languages used on social media, Tamil, a Dravidian language spoken by over 80 million people worldwide has seen a rise in abusive text targeting women (Chakravarthi et al., 2021). This phenomenon not only perpetuates gender-based discrimination but also poses significant challenges for natural language processing (NLP) systems tasked with detecting and reducing such content (Rajiakodi et al., 2025). The detection of abusive language in Tamil is particularly complex due to the language's rich morphology, code-mixing with English and other languages (Priyadharshini et al., 2022). Moreover cultural and contextual nuances often make it difficult for automated systems to accurately identify abusive content without misclassifying neutral text (Shanmugavadivel et al., 2022).

In our participation on Abusive Tamil and Malay alam Text Targeting Women on Social Media at DravidianLangTech@NAACL 2025 (Priyadharshini et al., 2023), we explored different models to detect abusive comments targeting women and addressed this problem with two significant contributions.

- Investigated the performance of various ML, DL and transformer-based models for detecting abusive comments.
- In particular leveraged the transformer-based XLM-RoBERTa model which demonstrated strong performance for abusive language detection in Tamil text.

This research shows that advanced models such as transformers can improve the detection of offensive comments in Tamil language. For more details, our code is available at https://github.com/MHD094/Abusive-Tamil.

2 Related Work

The detection of abusive language on social media particularly in low-resource languages like Tamil remains a critical yet underexplored area within NLP. While significant progress has been made in high-resource languages such as English, Tamil presents unique challenges including its rich morphology, informal writing styles and the prevalence

of code-mixing. Early research in abusive language detection focused on rule-based and machine learning methods using lexical features such as n-grams and part-of-speech tags (Rajalakshmi et al., 2022). With the advent of deep learning models and transformer-based architectures such as BERT (Devlin et al., 2019) have significantly improved abusive language detection.

In the domain of abusive language detection, (Ghanghor et al., 2021) presented a study on identifying offensive language and classifying memes in Dravidian languages particularly Tamil, Malayalam and Kannada. Gender-targeted abuse influenced by cultural nuances in Tamil remains a significant challenge. (Gong et al., 2021) tackle heterogeneous abusive language by introducing a YouTube dataset with sentence-level annotations. They propose a supervised attention model with multi-task learning, improving nuanced abuse detection. Their approach highlights the need for finer-grained annotations, relevant to Tamil abuse detection. (Mohan et al., 2025) introduced the Multimodal Tamil Hate (MATH) dataset, categorizing hate speech into offensive, sexist, racist and casteist types. This study emphasizes the need for culturally informed approaches to improve hate speech detection in Tamil.

The DravidianLangTech shared tasks (Chakravarthi et al., 2022) have further advanced research on NLP for Tamil by providing datasets for offensive language detection and sentiment analysis. These tasks have been instrumental in addressing challenges specific to Dravidian languages such as code-mixing and the compound nature of the languages. (Shanmugavadivel et al., 2022) demonstrated the effectiveness of machine learning for the sentiment analysis in Tamil code-mixed data. In spite of these advancements, challenges like the lack of large annotated datasets and the informal nature of social media writing persist. (Vetagiri et al., 2024) highlighted the need for collaborative efforts to overcome these barriers and improve abusive language detection in Tamil and other low-resource languages.

3 Task and Dataset Description

Abusive language targeting women has become a significant issue with the rise of social media, often reflecting societal biases and gender imbalances. This shared task focuses on abusive text detection in Tamil comments. It aims to identify whether a given comment contains abusive language directed at women or not. The dataset (Priyadharshini et al., 2023) and also (Priyadharshini et al., 2022) for this task comprises social media comments collected from YouTube discussions on controversial and sensitive topics where gender-based abuse is prevalent. This dataset supports accurate classification of abusive content into two binary classes as outlined below:

Abusive: Comments containing harmful or offensive language.

Non-Abusive: Comments free of offensive or abusive content.

Here, Table 1 provides the distribution of samples across training, validation and test sets. The dataset

Classes	Train	Valid	Test
Abusive	1366	278	305
Non-Abusive	1424	320	293
Total	2,790	598	598

Table 1: Dataset distribution.

is almost balanced with the Abusive class having 1,949 samples compared to 2,037 samples for the Non-Abusive class. The total dataset comprises 3,986 text which divided into training (2,790), validation (598) and test (598) sets.

4 Methodology

This section provides a concise summary of the methods and approaches adopted to address the problem outlined earlier. After thorough analysis, the transformer-based model XLM-RoBERTa demonstrates superior performance in our task. Figure 1 shows a visual representation of our methodology, highlighting the essential steps in the proposed approach.

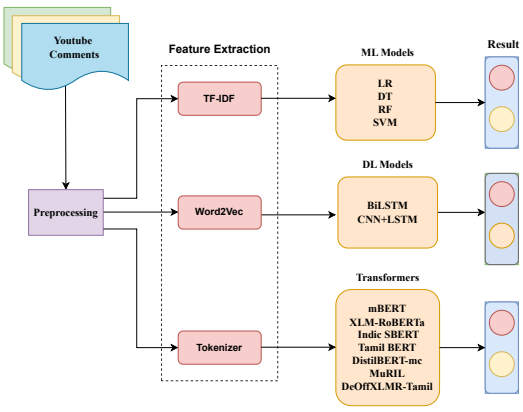


Figure 1: An abstract view of the proposed methodology

4.1 Preprocessing

Basic preprocessing steps such as removing special characters, emojis, punctuation and extra spaces were applied to clean the text. Indic-transliteration (Kunchukuttan, 2020) library used to convert code-mixed Tamil text into standardized Tamil for linguistic consistency and model compatibility.

4.2 Feature Extraction

To capture various features for different model types, three feature extraction techniques were applied. Machine learning models employ Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988) to represent text features. Deep learning models utilize word embeddings generated through the Word2Vec approach (Mikolov et al., 2013) for richer semantic information. Transformer-based models employ specialized tokenizers compatible with their architectures to efficiently process input sequences.

4.3 Model Building

In our research, we examined several ML, DL and transformer-based models.

4.3.1 ML models

We trained traditional ML models such as Logistic Regression (LR), Decision Trees (DT), Random Forest (RF) and Support Vector Machines (SVM) on feature representations like TF-IDF. These models rely on statistical patterns but may face challenges in understanding complex contextual relationships in code-mixed Tamil text.

4.3.2 DL models

The deep learning models include BiLSTM and a hybrid CNN+LSTM model. These models capture the semantic structure of the text and dependencies in code-mixed Tamil using pre-trained word embeddings. Each DL model was trained for 5 epochs with a batch size of 32.

4.3.3 Transformer-based models

The transformer-based models include mBERT (Ram et al., 2024), XLM-RoBERTa (Conneau et al., 2020), Indic SBERT (Farsi et al., 2024), Tamil BERT (Raihan et al., 2024), DistilBERT-mc (Rajalakshmi et al., 2023), MuRIL (Khanuja et al., 2021), and DeOffXLMR-Tamil. Fine-tuned on our dataset with transformer-specific tokenizers, these models excel at capturing long-range dependencies and context. They improve accuracy in Tamil abusive language detection by utilizing pre-trained knowledge from multilingual datasets, handling regional dialects and code-switching effectively.

5 Results & Discussion

This section presents a comparative analysis of the performance achieved by various machine learning, deep learning and transformer-based methods for detecting abusive comments in Tamil. The evaluation highlights the effectiveness of different classifiers in predicting abusive content. Additionally, m-BERT and XLM-RoBERTa were fine-tuned by optimizing learning rates, batch sizes and epochs while maintaining the AdamW optimizer as summarized in Table 2.

Hyperparameters	**m-BERT**		**XLM-RoBERTa**	
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	1e-05	2e-05	3e-05	1e-05
Epochs	12	8	8	12
Batch size	32	16	16	32

Table 2: Summary of optimized hyperparameters

We fine-tuned hyperparameters including learning rates, batch sizes and epochs to improve model performance. Table 3 presents precision (P), recall (R) and macro-F1 (MF1) scores on the test dataset. Among machine learning models, Logistic Regression (LR) performed best with an MF1 of 0.71 followed by SVM (0.69). In deep learning, BiLSTM and CNN+LSTM scored 0.45 and 0.33 respectively. Transformer models outperformed all with XLM-RoBERTa achieving the highest MF1 of 0.80 followed by m-BERT, MuRIL and Indic SBERT (0.77). Tamil BERT and DistilBERT-mc showed competitive performance. XLM-RoBERTa was the most effective for Tamil abusive comment detection.

Classifier	**P**	**R**	**MF1**
LR	0.71	0.74	0.71
DT	0.61	0.61	0.61
RF	0.67	0.66	0.66
SVM	0.69	0.69	0.69
BiLSTM	0.55	0.52	0.45
CNN + LSTM	0.24	0.49	0.33
mBERT	0.77	0.77	0.77
XLM-RoBERTa	0.80	0.80	0.80
Indic SBERT	0.77	0.76	0.76
Tamil BERT	0.67	0.68	0.68
DistilBERT-mc	0.75	0.74	0.74
MuRIL	0.77	0.77	0.77
DeOffXLMR-Tamil	0.76	0.76	0.76

Table 3: Results of several models on the test dataset

5.1 Quantitative Discussion

Figure 2 represents the confusion matrix for our XLM-RoBERTa model. The results highlight the effectiveness of transformer-based architectures especially XLM-RoBERTa in identifying abusive Tamil text. The model effectively classifies 245 Non-Abusive (label-0) and 234 Abusive (label-1) instances which demonstrates strong performance. However some misclassifications occur, 48 label-0 samples are predicted as label-1 and 71 label-1 samples are incorrectly classified as label-0. These results highlight the model's overall effectiveness.

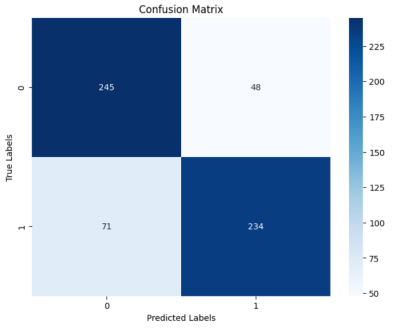


Figure 2: Confusion matrix of best performing model

5.2 Qualitative Discussion

Figure 3 displays sample predictions from our XLM-RoBERTa model. Samples 2, 3, 5, 6 and 8 are correctly classified which demonstrates the model's ability to process diverse linguistic constructs in Tamil text. However, challenges remain with samples 1, 4, and 7 which are misclassified as 0 instead of 1 due to implicit abuse, sarcasm and contextual complexity in Tamil language.

Sample Text	Actual	Predicted
சுகந்தி பொய் சொல்லுறா மோடம் பொய் சொல்லுறா (Is Sugandhi lying? Is Modem lying?)	1	0
மேடம், இப்போ பேசறதையும் நம்பாதீங்க, இதுவும் கண்டண்ட் தான் இருக்கும் (Madam, don't believe what I'm saying now, this too will be a conspiracy.)	0	0
என் வாழ்வில் இப்படி ஒரு கட்சியை பார்த்து பார்த்து சுமார் 30 தடவை பார்த்து பார்த்து சிறிது சிறிது ருசித்து சிரிப்பு வந்து விட்டது. சகோதரி லட்சுமி மிகவும் அருமையாக பேட்டி காண்பது சிரிப்பு வருது. இந்த திவ்யா ஒரு கோமாளி பெண். சுகந்தி பேச்சில் உண்மை தெரியுது. சகோதரி லட்சுமி நல்ல திரமையான ஒரு தொகுப்பாளர். கார்த்திக் எங்கே எங்கே எங்கே...பாவம் சுகந்தி (I have watched such a party in my life about 30 times, tasting it little by little, and it made me laugh. Sister Lakshmi gives an excellent interview, which is funny to watch. This Divya is a complete clown. The truth is evident in Sugandhi's speech. Sister Lakshmi is a very talented host. Karthik, where are you, where are you, where are you... Poor Sugandhi!)	0	0
பேய்க்கு பேய்க்கும் சண்ட அத ஊரே வேடிக்க பாக்குது (The city is a place where ghosts and ghosts are hunted.)	1	0
இப்படி இவங்க இல்லனா உங்களுக்கு வேலை இல்லை (Without these people, you wouldn't have a job.)	0	0
Divya எதுக்கு சுகன்யா எதிர்க்க என்ன காரணம் பாலா வுக்காக பாலா யாரு மதனுடய தம்பி...mathan யாரு அவரு mari Selva குழந்தை ய தப்ப பேசுனா...mariselva யாரு அவரு என்ன திட்டுநாரு...evlovo பிரச்சனை பிரச்சனை இருக்கு நாட்டுல Laxmi mam Enna ithu.... (Divya, why is Sukanya opposing Bala? Bala is the brother of Mathan...who is Mathan? Mari Selva, is the child talking nonsense...who is Mariselva? What is he scolding...evlovo, there is a problem in the country, Laxmi mam, Enna ithu....)	0	0
வாவ் சூப்பர் தூக்கி போட்டு மிதிக்க கூட்டங்களே அதே பெரிய விஷயம் (Wow, super, the crowds are the same big thing.)	1	0
இந்த பெண்ணுக்கு தரமான முடிவு அந்த ஊர் ஆம்பளங்க கையில இருக்கு. (The quality of this woman's life is in the hands of the town's mayor.)	1	1

Figure 3: Examples of the XLM-RoBERTa model predicted outputs

6 Conclusion

This research provides a comprehensive comparison of various machine learning, deep learning and transformer-based methods for detecting abusive comments in Tamil. The evaluation demonstrated that transformer models especially XLM-RoBERTa outperformed all other methods by achieving the highest macro-F1 score of 0.80 followed by m-BERT, MuRIL and Indic SBERT(.77). Among machine learning models, Logistic Regression showed the best performance with an MF1 of 0.71. Despite these advancements, there remain several areas for future research. Expanding datasets to include more diversity will improve model generalizability. Additionally integrating multimodal data such as images and videos could further enhance the detection of abusive content. Lastly, future research should focus on exploring the sociocultural factors driving gender-targeted abuse and developing interventions to address these issues at their core.

Limitations

One of the key limitations of this work stems from the challenges involved in preprocessing code-mixed Tamil text. The conversion of mixed language content through transliteration techniques may not fully capture all cultural context inherent in Tamil comments. This approach can lead to inaccuracies in identifying abusive language as the model may struggle with code-switching or non-standard expressions used in these comments. Furthermore the lack of sufficient high-quality annotated data for abusive language detection in low-resource languages such as Tamil restricts the model's effectiveness. The presence of implicit bias and regional dialect variations within the Tamil language adds another layer of complexity that could impact performance.

References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Cn, Sangeetha S, Malliga Subramanian, Kogilavani Shanmugavadivel, Parameswari Krishnamurthy, Adeep Hande, Siddhanth U Hegde, Roshan Nayak, and Swetha Valli. 2022. Findings of the shared task on multi-task learning in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 286–291, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl,

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Salman Farsi, Asrarul Eusha, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech EACL2024: Hate and offensive language detection in Telugu code-mixed text using sentence similarity BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 193–199, St. Julian's, Malta. Association for Computational Linguistics.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

Hongyu Gong, Alberto Valido, Katherine M. Ingram, Giulia Fanti, Suma Bhat, and Dorothy L. Espelage. 2021. Abusive language detection in heterogeneous contexts: Dataset collection and the role of supervised attention. *Preprint*, arXiv:2105.11119.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vishnu Subramanian, and Partha Pratim Talukdar. 2021. Muril: Multilingual representations for indian languages. *ArXiv*, abs/2103.10730.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Jayanth Mohan, Spandana Reddy Mekapati, Premjith B, Jyothish Lal G, and Bharathi Raja Chakravarthi. 2025. A multimodal approach for hate and offensive content detection in tamil: From corpus creation to model development. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Kathiravan Pannerselvam, Saranya Rajiakodi, Rahul Ponnusamy, and Sajeetha Thavareesan. 2023. CSS-CUTN@DravidianLangTech:abusive comments detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 306–312, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Abu Raihan, Tanzim Rahman, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@StressIdent_LT-EDI@EACL2024: Stress identification using Tamil-Telugu BERT. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 265–270, St. Julian's, Malta. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022. Dlrg@dravidianlangtech-acl2022: Abusive comment detection in tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Faerie Mattins R., Pavitra Vasudevan, and Anand Kumar M. 2023. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech Language*, 78:101464.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

N Prabhu Ram, T Meeradevi, P Sendhuraharish, S Yogesh, and C VasanthaKumar. 2024. Multi-class emotion classification on tamil and tulu code-mixed text. In *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 231–236. IEEE.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech Language*, 76:101407.

Advaitha Vetagiri, Gyandeep Kalita, Eisha Halder, Chetna Taparia, Partha Pakray, and Riyanka Manna. 2024. Breaking the silence detecting and mitigating gendered abuse in hindi, tamil, and indian english online spaces. *Preprint*, arXiv:2404.02013.

CUET_Absolute_Zero@DravidianLangTech 2025: Detecting AI-Generated Product Reviews in Malayalam and Tamil Language Using Transformer Models

Anindo Barua, Sidratul Muntaha, Momtazul Arefin Labib, Samia Rahman, Udoy Das[†], Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
[†]East Delta University, Bangladesh
{u2004040, u2004041, u1904111, u1904022}@student.cuet.ac.bd,
udoy.d@eastdelta.edu.bd, hasanmurad@cuet.ac.bd

Abstract

Artificial Intelligence (AI) is opening new doors of learning and interaction. However, it has its share of problems. One major issue is the ability of AI to generate text that resembles human-written text. So, how can we tell apart human-written text from AI-generated text? With this in mind, we have worked on detecting AI-generated product reviews in Dravidian languages, mainly Malayalam and Tamil. The "Shared Task on Detecting AI-Generated Product Reviews in Dravidian Languages," held as part of the DravidianLangTech Workshop at NAACL 2025 has provided a dataset categorized into two categories, human-written review and AI-generated review. We have implemented four machine learning models (Random Forest, Support Vector Machine, Decision Tree, and XGBoost), four deep learning models (Long Short-Term Memory, Bidirectional Long Short-Term Memory, Gated Recurrent Unit, and Recurrent Neural Network), and three fine-tuned transformer-based on their performance in detecting AI-generated text (AI-Human-Detector, Detect-AI-Text, and E5-Small-Lora-AI-Generated-Detector). We have conducted a comparative study among all the models by training and evaluating each model on the dataset. We have discovered that the transformer, E5-Small-Lora-AI-Generated-Detector, has provided the best result with an F1 score of 0.8994 on the test set ranking 7th in the Malayalam language. Tamil has a higher token overlap and richer morphology than Malayalam. Thus, we obtained a worse F1 score of 0.5877 ranking 28th position in the Tamil language among all participants in the shared task.

1 Introduction

AI-generated content has created a significant challenge in distinguishing authenticity. The misuse of AI can even lead to the spread of misinformation. For example, online product reviews that influence consumer decision-making are now raising concerns about their trustworthiness. A significant amount of previous studies have been conducted on AI-generated product review detection. The majority of works have been done in high-resource languages, like English (Mikros et al., 2023,Abburi et al., 2023,Valdez-Valenzuela et al., 2024,Marchitan et al., 2024). But, little has been done for low-resource languages Malayalam and Tamil. Difficult lexemes and no specific pattern for tokens make the detection of AI-generated Malayalam and Tamil product reviews difficult. Traditional Machine-Learning models have been found to struggle with the contextual awareness and linguistic complexities of Tamil and Malayalam (Islam et al., 2023). Likewise, Deep Learning-based RNN models such as GRU, LSTM, and BiLSTM have been found to struggle with capturing long-range dependencies and contextual nuances. (Gaggar et al., 2023)

Despite the demonstrated accuracy and widespread adoption in various natural language processing tasks, the Transformer-based approaches have not been utilized for AI-generated product review detection in Tamil and Malayalam language. This brings us to the primary objective of our paper, which is detecting AI-generated product reviews in Dravidian languages, mainly in Malayalam and Tamil, using transformer models. The "Shared Task on Detecting AI-generated Product Reviews in Dravidian Languages", at NAACL 2025 has provided a balanced yet limited dataset categorized into two categories, human-written reviews and AI-generated reviews.

We have implemented four machine learning models (Random Forest, SVM, Decision Tree, and XGboost), four deep learning models (RNN, GRU, LSTM, and BilSTM), and three transformer-based models (Ai-Human-Detector, Detect-Ai-Text, and E5-Small-Lora-Ai-Generated-Detector). We have conducted

a comparative study among the models by evaluating each model on the dataset.

The transformer-based (Alshammari et al., 2024, Mo et al., 2024) approaches that rely on self-attention mechanisms can capture long-range dependencies along with contextual connections within texts. Thus, languages like Malayalam and Tamil, having complex morphological and syntactic structures, are generalized well by transformer-based models rather than machine learning and deep learning models. Among the three fine-tuned transformer-based models, we have found that the transformer "E5-Small-Lora-Ai-Generated-Detector" has provided the best result. In the case of the Malayalam language, it has obtained an F1 score of 0.8994 on the test set, ranking 7th. For the Tamil language, due to a higher token overlap and richer morphology, it has ranked 28th with an F1 score of 0.5877 among all participants in the shared task (Premjith et al., 2025). The core contributions of this research work are:

- To implement and compare the traditional ML models, deep learning models, and transformer-based models.
- To handle insufficient data by augmentation, to conduct a detailed error analysis, and to investigate the causes of misclassification.

The implementation details have been provided in the following: GitHub Repository.

2 Related Work

The previous studies on AI-generated Text Detection can be categorized under Machine Learning, Deep Learning, and Transformer-Based approaches.

Traditional Machine Learning (ML) techniques have been applied for AI-generated Text Detection in online social media platforms (Gaggar et al., 2023). Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) have been used widely for AI-generated text detection with SVM providing the best result (Cingillioglu, 2023).

In comparison, Deep learning-based approaches are less dependent on explicitly defined features as they learn patterns and features automatically. The models integrate various layers, including LSTM, Transformer, and CNN, to perform tasks such as text classification and sequence labeling. This combination allows the model to effectively capture linguistic patterns improving text detection capabilities (Mo et al., 2024).

Generative language models like BERT, RoBERTa, and GPT have been used in detecting AI-based techniques (Mikros et al., 2023). DistilBert-Base-Uncased Model, Detect-Ai-Text, And E5-Small-Lora-Ai-Generated-Detector have demonstrated their effectiveness in the field of AI-generated text detection. In addition, several shared tasks (Nguyen et al., 2023,Maloyan et al., 2022) are accelerating the research effort leading to greater refinement of the AI-generated text detection methods.

3 Dataset

In the shared task "Detecting AI-Generated Product Reviews in Dravidian Languages" at the DravidianLangTech Workshop at NAACL 2025, the dataset provided for AI-generated product reviews contains Malayalam and Tamil language reviews. Table 1 and Table 2 contain the Tamil and Malayalam train datasets respectively. Initially, the datasets are limited in size. We have used data augmentation via back translation using the substitution method to compensate for data scarcity. Post-augmentation, the Tamil training split expanded to 806 human-written and 810 AI-generated reviews, while the Malayalam split expanded to 800 of each, providing more data for improved model training.

Table 1: Category-wise distribution in the Tamil dataset

Sets	AI	HUMAN	Total
Train	405	405	810
Development	405	401	806
Test	50	50	100
Total	860	856	1716

Table 2: Category-wise distribution in the Malayalam dataset

Sets	AI	HUMAN	Total
Train	400	400	800
Development	400	400	800
Test	100	100	200
Total	900	900	1800

4 Methodology

We have worked on a binary classification task involving low-resource languages to classify product reviews as AI-generated or human-written. At the outset, feature extraction has been performed. Multiple ML and DL algorithms were applied for analysis.

ML-based approaches, including Decision Tree, Support Vector Machine(SVM), Random Forest, and XGBoost have been used. SVM has been incorporated with a soft margin in the hyperplane. Deep Learning-based approaches, including RNN, LSTM, GRU, and BiLSTM have been used. This dataset is tokenized using NLTK's word-level tokenizer, which outputs a list of individual words and punctuation marks. We have applied Word2Vec for feature extraction due to its simplicity and ease of implementation. We could have replaced it with FastText, which handles out-of-vocabulary words better by leveraging subword information. Yet, we focused on transformer-based models for their superior performance in capturing complex linguistic patterns and the time constraints of the shared task. This decision not to incorporate FastText embeddings represents a limitation of our study.

Additionally, the system has been enhanced using different transformer models, as illustrated in Figure 1.

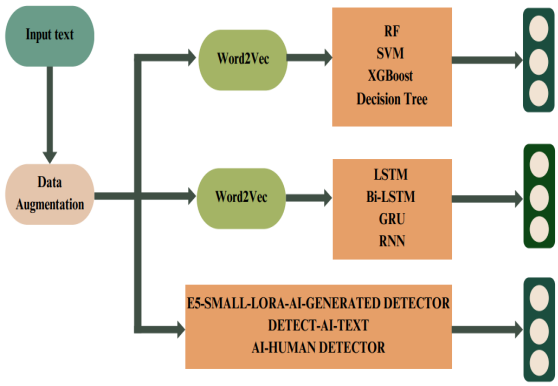


Figure 1: Abstract process of AI-generated product review detection

Three fine-tuned transformer-based models, E5-Small-LoRA-AI-Generated-Detector, AI-Human-Detector, and Detect-AI-Text, have been fine-tuned on back-translated Tamil and Malayalam text using the Trainer API of HuggingFace. Due to the limited data available in the dataset, we have implemented Synonym-based Augmentation using Contextual embedding (Pavlyshenko and Stasiuk, 2023). At first, we have used the Google-trans library where the text data has been back-translated through English to introduce variations in the dataset. Additional augmentation has been performed using the ContextualWordEmbsAug class of the nlpaug library, which optimizes models like bert-base-uncased for word substitutions based on context. Underrepresented labels have been identified, and new samples have been generated to balance the dataset by applying augmentation to randomly selected rows. We have aimed to expand the size of the dataset and thus our training dataset has increased from 800 entries to 1600 for Malayalam and 808 entries to 1616 for Tamil.

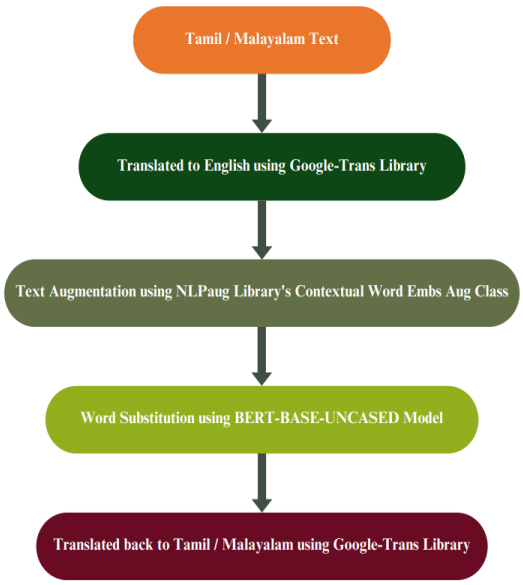


Figure 2: Data Augmentation technique

5 Results and Analysis

Our performance comparison among the ML, DL, and transformer-based approaches is shown in this section.

5.1 Parameter Setting

In Table 3, lr, optim, bs, wd, and wr represent learning_rate, optimizer, batch_size, weight_decay, and warmup_ratio respectively. Model name AHD represents AI-Human-Detector, E5-SLAGD represents E5–Small-Lora-Ai-Generated-Detector, and DAT represents Detect-Ai-Text.

Table 3: Parameter settings for different models

Model	lr	optim	bs	wd	wr
AHD	$2e^{-5}$	AdamW	4	0.01	0.1
e5-SLAGD	$2e^{-5}$	AdamW	4	0.01	0.1
DAT	$2e^{-5}$	AdamW	4	0.01	0.1
LSTM	$1e^{-3}$	Adam	32	-	-
BiLSTM	$1e^{-3}$	Adam	32	-	-
GRU	$1e^{-3}$	Adam	32	-	-
RNN	$1e^{-3}$	Adam	32	-	-

5.2 Evaluation Metrics

The performance of various models has been evaluated by calculating the precision (P), recall (R), and F1-Score on the test set.

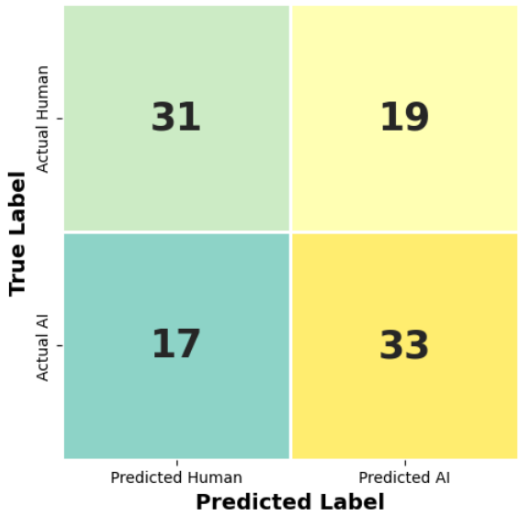
5.3 Comparative Analysis

Table 4: Performance of different systems on Malayalam and Tamil test datasets

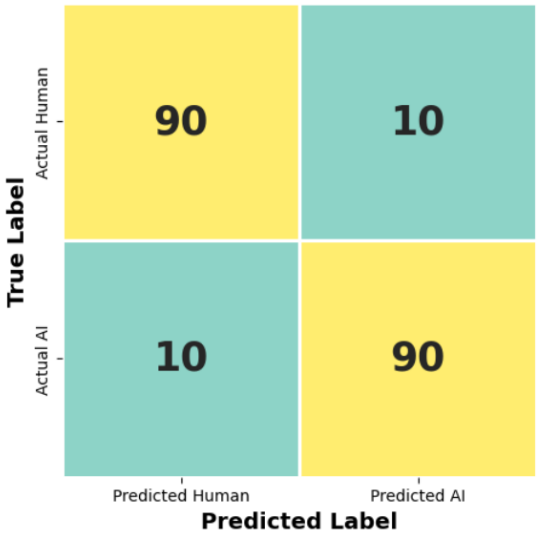
Classifier	Malayalam			Tamil		
	P	R	F1	P	R	F1
ML						
DT	0.15	0.09	0.11	0.15	0.09	0.13
RF	0.36	0.23	0.39	0.16	0.13	0.43
SVM	0.21	0.13	0.23	0.16	0.11	0.12
XGBOOST	0.32	0.11	0.35	0.21	0.13	0.35
DL						
BiLSTM	0.25	0.25	0.33	0.23	0.48	0.31
LSTM	0.25	0.50	0.33	0.23	0.48	0.31
GRU	0.77	0.53	0.49	0.46	0.51	0.38
RNN	0.43	0.43	0.37	0.27	0.52	0.36
TF						
E5-SLAGD	**0.91**	**0.90**	**0.90**	**0.65**	**0.62**	**0.59**
AHD	0.25	0.50	0.33	0.70	0.70	0.48
DAT	0.87	0.87	0.87	0.26	0.48	0.33

We have conducted a comparative study among four ML models Random Forest, Support Vector Machine, Decision Tree, and XGBoost, four DL models (LSTM, BiLSTM, GRU, and RNN), and three transformer-based models (AI-Human-Detector, Detect-AI-Text, and E5-Small-LoRA-AI-Generated-Detector). Among the ML models, XGBoost achieved the highest F1 score. DL model GRU has outperformed ML models. Transformer models have significantly outperformed both ML and DL models. The E5-Small-LoRA-AI-Generated-Detector (E5-SLAGD) model, with an F1 score of 0.8994 for Malayalam and 0.592 for Tamil, has achieved the best results ranking 7th in Malayalam and 28th in Tamil among all participants in the shared task.

5.4 Error Analysis



(a) Confusion Matrix for Malayalam



(b) Confusion Matrix for Tamil

Figure 3: Confusion Matrices for Malayalam and Tamil

Table 4 shows that E5-Small-Lora-AI-Generated-Detector did better than all others. We have created confusion matrices shown in Figure 3a and 3b) for a better understanding of the system. The False Positive (FP) for Malayalam was 4%, while FP for Tamil was 5% for E5.

AI-generated reviews often mimic human speech patterns, making misclassification a central issue. Tamil is morphologically richer than Malayalam and the overlapping of tokens is more.

Thus, the model faced greater challenges with Tamil compared to Malayalam.

6 Limitation

While NLTKTokenizer provided a baseline, our future work will investigate subword tokenization techniques such as Byte-Pair Encoding (BPE) or WordPiece, to better handle morphologically complex words. Also, the specific hyperparameter values (learning rate, optimizer, batch size, weight decay, and warmup ratio) were initially selected based on recommendations from the original model publications. Due to time constraints, a formal hyperparameter optimization strategy, such as grid search or Bayesian optimization, could not be employed. Our future works will explore more systematic hyperparameter tuning methods to potentially improve model performance.

7 Conclusion

In this paper, a comparative study has been carried out among various machine learning, deep learning, and transformer-based models for detecting AI-generated product reviews in Malayalam and Tamil languages. We have utilized the dataset of the "Shared Task on Detecting AI-Generated Product Reviews in Dravidian Languages" at NAACL for training the models. Three pre-trained models were used among which, AI-Human-Detector has outperformed other models with an F1 score of 0.8994 for Malayalam and 0.5877 for Tamil. We have calculated the Jaccard Index for both languages to quantify token overlap between AI-generated and human-written reviews. Our average Jaccard Index for Tamil at 0.35 is greater than that for Malayalam at 0.22. This suggests the token overlap in Tamil is higher where AI generated and human written have a higher percentage of tokens. It is likely that this larger overlap is why the two classes are more difficult to distinguish in Tamil. The overlap between the target tokens in the dataset has caused misclassification despite the Transformer-based models having excellent contextual understanding. The tokenization of Tamil is more complicated due to richer morphology for which the Tamil model performed worse than the Malayalam model. To tackle these issues in the near future, we plan to explore advanced transformer architectures and enhanced data augmentation.

Ethical Statement

The data analysis and model development tools and technologies used for this study are used ethically and responsibly. The purpose of our work is to create a system that detects AI product reviews for the good of maintaining transparency and authenticity on online platforms. We believe knowledge is for sharing, so we will share our work and contribute to the development of AI-generated content detection in low-resource languages such as Malayalam and Tamil.

References

Harika Abburi, Kalyani Roy, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. A simple yet efficient ensemble approach for ai-generated text detection. *Preprint*, arXiv:2311.03084.

Hamed Alshammari, Ahmed El-Sayed, and Khaled Elleithy. 2024. Ai-generated text detector for arabic language using encoder-based transformer architecture. *Big Data and Cognitive Computing*, 8(3).

Ilker Cingillioglu. 2023. Detecting ai-generated essays: the chatgpt challenge. *The International Journal of Information and Learning Technology*, 40(3):259–268.

Raghav Gaggar, Ashish Bhagchandani, and Harsh Oza. 2023. Machine-generated text detection using deep learning. *Preprint*, arXiv:2311.15425.

Niful Islam, Debopom Sutradhar, Humaira Noor, Jarin Tasnim Raya, Monowara Tabassum Maisha, and Dewan Md Farid. 2023. Distinguishing human generated text from chatgpt generated text using machine learning. *Preprint*, arXiv:2306.01761.

Narek Maloyan, Bulat Nutfullin, and Eugene Ilyushin. 2022. Dialog-22 ruatd generated text detection. *ArXiv*, abs/2206.08029.

Teodor-George Marchitan, Claudiu Creanga, and Liviu P. Dinu. 2024. Transformer and hybrid deep learning based models for machine-generated text detection. *Preprint*, arXiv:2405.17964.

George K. Mikros, Athanasios Koursaris, Dimitrios Bilianos, and George Markopoulos. 2023. Ai-writing detection using an ensemble of transformers and stylometric features. In *IberLEF@SEPLN*.

Yuhong Mo, Hao Qin, Yushan Dong, Ziyi Zhu, and Zhenglin Li. 2024. Large language model (llm) ai text generation detection based on transformer deep learning algorithm. *ArXiv*, abs/2405.06652.

Duke Nguyen, Khaing Myat Noe Naing, and Aditya Joshi. 2023. Stacking the odds: Transformer-based

ensemble for ai-generated text detection. pages 173–178.

B. Pavlyshenko and Mykola Stasiuk. 2023. Augmentation in a binary text classification task. *2023 IEEE 13th International Conference on Electronics and Information Technologies (ELIT)*, pages 177–180.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Andric Valdez-Valenzuela, Ricardo Loth Zavala-Reyes, Victor Giovanni Morales Murillo, and Helena Gómez-Adorno. 2024. The iimasnlp team at iberautextification 2024: Integrating graph neural networks, multilingual llms, and stylometry for automatic text identification.

MNLP@DravidianLangTech 2025: Transformers vs. Traditional Machine Learning: Analyzing Sentiment in Tamil Social Media Posts

Abhay Vishwakamra
Department of CSE
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
vishwakarmaabhay10@gmail.com

Abhinav Kumar
Department of CSE
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
abhik@mnnit.ac.in

Abstract

Sentiment analysis in Natural Language Processing (NLP) aims to categorize opinions in text. In the political domain, understanding public sentiment is crucial for influencing policymaking. Social media platforms like X (Twitter) provide abundant sources of real-time political discourse. This study focuses on political multiclass sentiment analysis of Tamil comments from X, classifying sentiments into seven categories: substantiated, sarcastic, opinionated, positive, negative, neutral, and none of the above. A number of traditional machine learning such as Complement Naive Bayes, Voting Classifier (an ensemble of Decision Tree, SVM, Naive Bayes, K-Nearest Neighbors, and Logistic Regression) and deep learning models such as LSTM, deBERTa, and a hybrid approach combining deBERTa embeddings with an LSTM layer are implemented. The proposed ensemble-based voting classifier achieved best performance among all implemented models with an accuracy of 0.3750, precision of 0.3387, recall of 0.3250, and macro-F_1-score of 0.3227.

1 Introduction

Sentiment analysis, a key task in Natural Language Processing (NLP), involves categorizing opinions in text (Kumar et al., 2020; Mishra et al., 2021). In the political domain, understanding public sentiment is crucial for policymaking. Social media platforms like X (formerly Twitter) provide real-time political discourse, but analyzing sentiments, particularly in code-mixed languages, presents unique challenges (Kumar et al., 2021, 2023; Kumari and Kumar, 2021). Code-mixing (Bokamba, 1988), common in multilingual communities, involves switching between languages in a single text. In India, users often blend English with regional languages like Tamil, creating challenges for sentiment analysis. Tamil, with its rich literary heritage, is frequently written in Roman script on social media, resulting in code-mixed content that complicates NLP tasks.

(Thavareesan and Mahesan, 2021) applied K-Means and KNN for sentiment analysis in Tamil texts. (Shanmugavadivel et al., 2022) evaluated machine learning models for sentiment classification in Tamil code-mixed tweets. (Anbukkarasi and Varadhaganapathy, 2020) explored deep neural networks, particularly DBLSTM, highlighting their ability to capture complex linguistic patterns.

A shared task on political multiclass sentiment analysis of Tamil social media posts was introduced during the DravidianLangTech@NAACL 2025 workshop (Durairaj et al., 2025). This task involved categorizing sentiments into seven categories: opinionated, sarcastic, substantiated, positive, negative, neutral, and none of the above. There are several deep learning models like LSTM, deBERTa, and a hybrid approach that combines deBERTa embeddings with an LSTM layer, as well as several traditional machine learning models like Complement Naive Bayes, Voting Classifier (an ensemble of Decision Tree, SVM, Naive Bayes, K-Nearest Neighbours, and Logistic Regression), and others are implemented.

The rest of the paper is summarized as follows: Section 2 introduces the dataset, Section 3 explains the proposed model, and the outcome of the proposed model is listed in Section 4, the discussion of findings and conclusion are listed in Section 5, limitations of proposed model and future directions are listed in Section 6.

2 Data Description

The dataset used for this analysis is provided by the DravidianLangTech@NAACL 2025 shared task[1], which focuses on political sentiment analysis in

[1] https://codalab.lisn.upsaclay.fr/competitions/20702#learn_the_details-overview

Tamil. It consists of a collection of Tamil-language tweets gathered from X, capturing a broad spectrum of political discussions. Each tweet is annotated with one of the seven sentiment categories: (i) substantiated, (ii) sarcastic, (iii) opinionated, (iv) positive, (v) negative, (vi) neutral, and (vii) none of the above. The overall distribution of the dataset is listed in Table 1.

Table 1: Distribution of Labels in Training and Validation Sets

Label	Training	Validation
Opinionated	1,361	153
Sarcastic	790	115
Neutral	637	84
Positive	575	69
Substantiated	412	52
Negative	406	51
None of the above	171	20
Total	**3,352**	**544**

3 Methodology

To tackle class imbalance issue, we used Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) on our train dataset. Specifically, this method essentially generate artificial samples from linear combinations of two or more minority classes examples so that we can again have a more balanced sample.

Five different models were developed to identify hate or offensive contents in Dravidian posts; (i) Complement Naive Bayes, (ii) Voting Classifier, (iii) Long-Short Term-Memory (LSTM) , (iv) Transfer learning-based model, and (v) Hybrid model. In this section, we explain the working of each model in detail.

3.1 Complement Naive Bayes

We use the Complement Naive Bayes (Seref and Bostanci, 2019) (CNB) algorithm for text classification, ideal for imbalanced datasets as it adjusts weights using the complement of each class to reduce sensitivity to imbalances. Text data is preprocessed using Count Vectorizer, which converts text into a matrix of token counts representing word frequencies, serving as input for the classifier.

3.2 Voting Classifier

We use a Voting (Kuncheva and Rodríguez, 2014) Classifier with Decision Tree (Song and Ying, 2015), SVM (Jakkula, 2006), Multinomial Naive Bayes (Kibriya et al., 2005), K-Nearest Neighbors (Guo et al., 2003), and Logistic Regression (DeMaris, 1995). Soft voting (Cao et al., 2015) combines the predicted probabilities from each model for a balanced consensus. Text data is preprocessed using Count Vectorizer, which converts text into a matrix of token counts for training the ensemble.

3.3 Deep Learning Model

We built an LSTM (Aston Zhang, 2020) based model for text classification with two stacked (Wang et al., 2018) LSTM layers (64 and 32 units), dropout layers for regularization, and a softmax output for 7-class prediction. Text preprocessing included IndicNLP (Kunchukuttan) for tokenization and normalization, followed by padding for uniform input length.

3.4 Transfer Learning Model

We utilized a transfer learning approach with the DeBERTa (He et al., 2020) V3 base multilingual model for text classification into 7 categories. We use a DeBERTa model since its disentangled attention mechanism and absolute position embeddings yield better contextual embeddings and superior performance than traditional BERT-based models with common pooling methods especially in low-resource multilingual settings. The preprocessor was configured using *DebertaV3TextClassifierPreprocessor* with a sequence length of 64 and waterfall truncation then preprocessed input was fed into the *DebertaV3TextClassifier*, utilizing pre-trained embeddings for efficient predictions.

3.5 Hybrid Model

We use a hybrid model that combines DeBERTa V3 with LSTM (Rai et al., 2022) to utilize their complementary strengths. DeBERTa generates contextual embeddings, while LSTM captures sequential dependencies, enhancing performance for text classification tasks.

The input text is preprocessed using the DeBERTa preprocessor, which tokenizes and prepares data for the transformer. The DeBERTa V3 model outputs embeddings with a shape of batch_size, sequence_length, embedding_dim i.e.(32, 64, 768), which are fed into an LSTM layer. The first LSTM layer, with 128 units and second LSTM layer with 64 units. Finally, a dense classification layer with 7

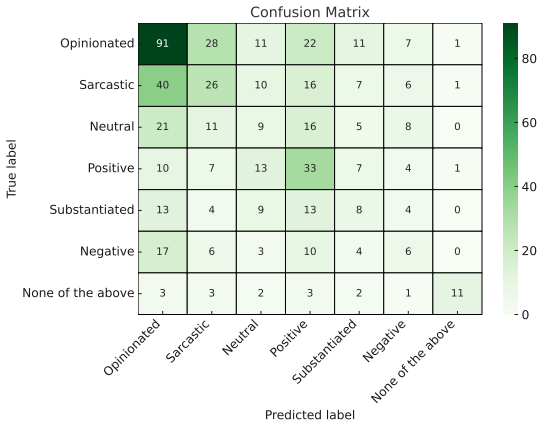


Figure 1: Confusion Matrix for Naive Bayes classifier

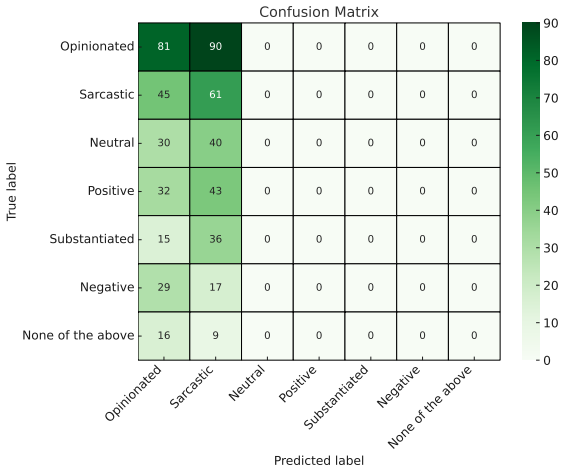


Figure 2: Confusion Matrix for DeBERTa model

units applies a softmax activation to produce class probabilities for predictions.

4 Result

This section has the results of the five models evaluated using accuracy, precision, recall, and macro-F_1-scores (Opitz and Burst, 2019). The results of different machine learning and deep learning models can be seen in Table 2. As can be seen in Table 2, the Voting Classifier outperforms the other models with an accuracy of 0.3750, precision of 0.3387, recall of 0.3250 and an macro-F_1-score of 0.3227. The Naive Bayes classifier shown good efficiency for text classification with an accuracy of 0.3382, precision of 0.3367, recall of 0.2962, and macro-F_1-score of 0.3059. Similar results were obtained by the hybrid model (DeBERTa + LSTM), which benefited from both contextual and sequential learning, with an accuracy of 0.3162, precision of 0.3143, recall of 0.2997, and macro-F_1-score of 0.3026.

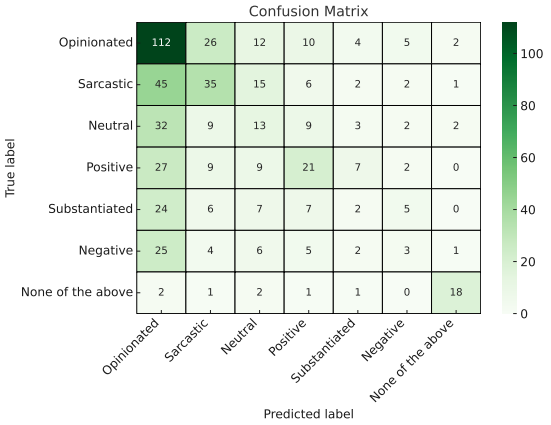


Figure 3: Confusion Matrix for Voting Classifier

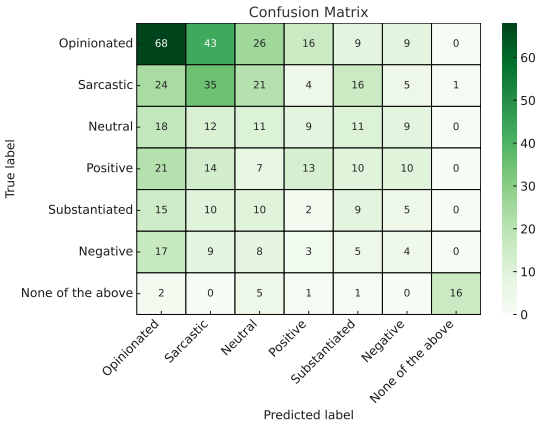


Figure 4: Confusion Matrix for LSTM model

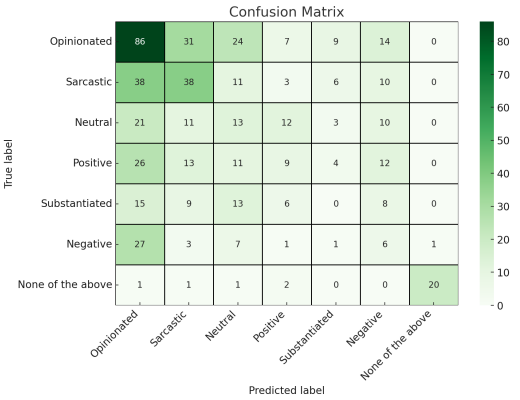


Figure 5: Confusion Matrix for Hybrid Model (DeBERTa with LSTM).

Table 2: Performance comparison of different models.

Model	Accuracy	Precision	Recall	macro-F_1-score
Naive Bayes	0.3382	0.3367	0.2962	0.3059
Voting Classifier	0.3750	0.3387	0.3250	0.3227
2-layer LSTM	0.2868	0.3252	0.2803	0.2964
DeBERTa_v3	0.2610	0.0761	0.1499	0.0986
DeBERTa Embeddings + LSTM layer	0.3162	0.3143	0.2997	0.3026

The 2-layer LSTM model showed moderate results, with an accuracy of 0.2868, precision of 0.3252, recall of 0.2803, and macro-F_1-score of 0.2964. The DeBERTa v3 model had the lowest metrics, with an accuracy of 0.2610, precision of 0.0761, recall of 0.1499, and macro-F_1-score of 0.0986, indicating that pre-trained embeddings alone might not fully capture the task's nuances. Overall, ensemble and hybrid models, like the Voting Classifier and DeBERTa embedding with LSTM, performed better compared with other implemented models.

5 Discussion and Conclusion

The confusion matrix (Heydarian et al., 2022) for Naive Bayes shows some misclassification (see Figure 1), particularly for Opinionated, Sarcastic, and Neutral classes. The Transfer Learning Model (DeBERTa) exhibits a high degree of misclassification (see Figure 2), especially for classes Opinionated, Sarcastic, and Neutral. The confusion matrix (see Figure 3) for the Voting Classifier shows a good balance of correct predictions and minimal misclassifications. The Deep Learning (2 stacked LSTM Layers) (see Figure 4) and Hybrid (DeBERTa embedding with LSTM) (see Figure 5) models show moderate performance with some misclassifications. The Hybrid model appears to have slightly better performance than the LSTM based on the confusion matrix.

The Voting Classifier performs best with the highest accuracy, precision, and recall. Naive Bayes and deBERTa show weaker results, while the LSTM and Hybrid models perform moderately. Overall, ensemble methods like the Voting Classifier are most effective for political sentiment analysis in Tamil tweets.

6 Limitations and Future Work

Since the oversampling of SMOTE might affect a model, class imbalance, continues to remain a challenge, as SMOTE generates synthetic samples alike to the real-world scenario but does not necessarily denote the complexities of natural language. As a result, the dataset is large and domain-specific, and providing additional annotated samples for dominant classes in particular can enhance the robustness of the dataset. Finally, real-time deployment of DeBERTa incurs considerable computational costs given its transformer architecture, so future work may explore lower-cost architectures or distillation (Gou et al., 2021) approaches.

In order to overcome these limitations, we intend to extend the dataset, investigate multimodal approaches in addition to text, further fine-tune model efficiency, and evaluate domain transfer over various Tamil dialects along with variations in political discourse.

The code for the proposed framework is available at:
https://github.com/abhay-43/Deep-Learning-Approach-for-Analyzing-Sentiment-in-Tamil-Social-Media-Posts.git

References

S Anbukkarasi and S Varadhaganapathy. 2020. Analyzing sentiment in tamil tweets using deep neural network. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 449–453. IEEE.

Mu Li Alexander J. Smola Aston Zhang, Zachary C. Lipton. 2020. Dive into deep learning. `https://d2l.ai/chapter_recurrent-modern/lstm.html`. Accessed: 2024-11-26.

Eyamba G Bokamba. 1988. Code-mixing, language variation, and linguistic theory:: Evidence from bantu languages. *Lingua*, 76(1):21–62.

Jingjing Cao, Sam Kwong, Ran Wang, Xiaodong Li, Ke Li, and Xiangfei Kong. 2015. Class-specific soft voting based multiple extreme learning machines ensemble. *Neurocomputing*, 149:275–284.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Alfred DeMaris. 1995. A tutorial in logistic regression. *Journal of Marriage and the Family*, pages 956–968.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Mohammadreza Heydarian, Thomas E Doyle, and Reza Samavi. 2022. Mlcm: Multi-label confusion matrix. *IEEE Access*, 10:19083–19095.

Vikramaditya Jakkula. 2006. Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37(2.5):3.

Ashraf M Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. 2005. Multinomial naive bayes for text categorization revisited. In *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*, pages 488–499. Springer.

Abhinav Kumar, Sunil Saumya, and Jyoti Prakash Singh. 2020. NITP-AI-NLP@ Dravidian-CodeMix-FIRE2020: a hybrid cnn and bi-lstm network for sentiment analysis of Dravidian code-mixed social media posts. In *FIRE (Working Notes)*, pages 582–590.

Abhinav Kumar, Sunil Saumya, and Jyoti Prakash Singh. 2021. An ensemble-based model for sentiment analysis of Dravidian Code-Mixed social media posts. In *FIRE (Working Notes)*, pages 950–958.

Abhinav Kumar, Jyoti Prakash Singh, and Amit Kumar Singh. 2023. Explainable BERT-LSTM stacking for sentiment analysis of covid-19 vaccination. *IEEE Transactions on Computational Social Systems*.

Jyoti Kumari and Abhinav Kumar. 2021. A deep neural network-based model for the sentiment analysis of dravidian code-mixed social media posts. *management*, 5:6.

Ludmila I Kuncheva and Juan J Rodríguez. 2014. A weighted voting framework for classifiers ensembles. *Knowledge and information systems*, 38:259–275.

Anoop Kunchukuttan. Indic nlp library. https://anoopkunchukuttan.github.io/indic_nlp_library/.

Ankit Kumar Mishra, Sunil Saumya, and Abhinav Kumar. 2021. Sentiment analysis of Dravidian-CodeMix language. In *FIRE (Working Notes)*, pages 1011–1019.

Juri Opitz and Sebastian Burst. 2019. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. 2022. Fake news classification using transformer based enhanced lstm and bert. *International Journal of Cognitive Computing in Engineering*, 3:98–105.

Berna Seref and Erkan Bostanci. 2019. Performance comparison of naïve bayes and complement naïve bayes algorithms. In *2019 6th international conference on electrical and electronics engineering (ICEEE)*, pages 131–138. IEEE.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech & Language*, 76:101407.

Yan-Yan Song and LU Ying. 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53. IEEE.

Jin Wang, Bo Peng, and Xuejie Zhang. 2018. Using a stacked residual lstm model for sentiment intensity prediction. *Neurocomputing*, 322:93–101.

shimig@DravidianLangTech2025: Stratification of Abusive content on Women in Social Media

Gersome Shimi
Madras Christian College,
Chennai, India
gshimi2022@gmail.com

C.Jerin Mahibha
Meenakshi Sundararajan
Engineering College,
Chennai, India
jerinmahibha@msec.edu.in

Durairaj Thenmozhi
Sri Sivasubramaniya Nadar
College of Engineering,
Chennai, India
theni_d@ssn.edu.in

Abstract

The social network is a trending medium for interaction and sharing content globally. The content is sensitive since it can create an impact and change the trends of stakeholder's thought as well as behavior. When the content is targeted towards women, it may be abusive or non-abusive and the identification is a tedious task. The content posted on social networks can be in English, code mix, or any low-resource language. The shared task Abusive Tamil and Malayalam Text targeting Women on Social Media was conducted as part of DravidianLangTech@NAACL 2025 organized by DravidianLangTech. The task is to identify the content given in Tamil or Malayalam or code mix as abusive or non-abusive. The task is accomplished for the South Indian languages Tamil and Malayalam using pretrained transformer model, BERT base multilingual cased and achieved the accuracy measure of 0.765 and 0.677.

1 Introduction

According to Statista, a Statistics portal for market data, market research, and market studies, the number of Internet users in the year 2024 is 5.44 billion. It is one third of the world's population and the number of YouTube users is approximately 2504 million as of April 2024[1]. People use social media platforms such as YouTube, Twitter, Instagram, Reddit, and Facebook to share their opinions, beliefs, and interests in all the state of affairs. It can be used positively in e-commerce, information transfer, advertisements, politics, hobbies, testimonies, education and training, recent happenings, etc. Alternatively, it can lead to the spread of hate speech and on-line harassment, which is termed cyberbully. It should be identified since it will cause psychological impact for the stakeholder even to depression.(Sari et al., 2022)

The trendy digital platforms, social media impact people at various levels, even the political and business scenario can be altered within next few hours in par with the comments posted. It can also target a particular individual or a group of individuals.(Priyadharshini et al., 2022) Hate speech and offensive language can harm various groups and can end with social problems, thus makes detection an essential task to reduce crime and promote harmony. Although significant research exists for languages like English, Dravidian languages such as Tamil and Malayalam lack focus.(Mahibha et al., 2021) When abusive words are aimed at gender, particularly on women it is defined as sexism. As it is one of the alarmed need of the social media, focusing on this issue, DravidianLangTech@NAACL2025 organized by DravidianLangTech initiated a shared task to identify abusive Tamil and Malayalam Text targeting Women on Social Media,

2 Related Work

Hate speech is a complicated and multifaceted issue that creates serious and widespread implications for human rights and the rule of law on democratic society. Addressing and preventing online hate speech presents particular challenges. The ongoing nature and effects of this issue have been recorded by the oversight bodies of the Council of Europe and various international organizations[2].

Social networks are rapidly expanding with different content in various low resource languages allows stakeholder to express their opinions with few restrictions. Most social media platforms allow users to share and express their thoughts to collect user comments and posts to offer channels of personalized interest. However, they are also used for negative actions, such as spreading ru-

[1] https://www.statista.com/topics/1145/internet-usage-worldwide

[2] https://www.coe.int/en/web/combating-hate-speech/what-is-hate-speech-and-why-is-it-a-problem-

mors and intimidating people with offensive words. Abusive language has attracted a lot of attention as social media platforms have become more popular.(Barman and Das, 2023)

2.1 Low Resource Languages

Low resource languages own complexity in terms of variation in writing and spoken style, unavailability of resources, and corpus. Words usage have different meanings when used in different communities. LLM (Large Language Model) is capable of grasping multiple languages and adapting to different contexts (Zhong et al., 2024). Low resource language, Kannada, Malayalam, Telugu, and Tamil, obtains less attention due to unavailability of the corpus. The ensemble transformer model is applied for the classification, obtain the f1 score of 0.66 and 0.72 for Kannada code mix and Malayalam, respectively. (Roy, 2024)

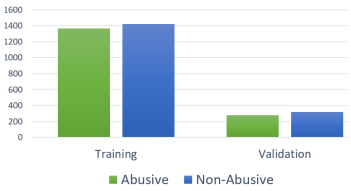


Figure 1: Distribution of Dataset-Tamil

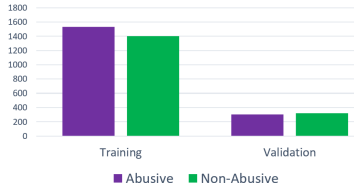


Figure 2: Distribution of Dataset-Malayalam

Language	Text	Label
Tamil	என்ன திமிர் இந்த பொண்ணுக்கு.....மக்களே இன்னும் இவளுடைய வீடியோக்களை பார்த்தால் உங்களை விட முட்டாள்கள் யாருமில்லை	Abusive
Tamil	You tube ல இப்படி எல்லாம் முன்னேற்றம் நம் சேனலையும வந்து முன்னேற்றுங்கள்!!	Non-Abusive
Malayalam	ചുമ്മാതല്ല ഇവളെ പണ്ട് ലവർ അലക്കി വിട്ടത്...വായിൽ കിടക്കുന്ന നാക്ക് അങ്ങനെ ആണ്	Abusive
Malayalam	ഒരു സിനിമയിൽ ജഗതി മദാമ്മയായിട്ട് വരുന്നുണ്ടല്ലോ. പെട്ടന്ന് അതോർമ്മ വന്നു	Non-Abusive

Figure 3: Sample Dataset

2.2 Preprocessing

Most instances are misidentified because of the presence of insignificant words. Preprocessing is an important process to feed the model with quality data in terms of size, improves the model performance.(Kumari, 2022) Text cleaning is the removal of insignificant words from the dataset sentence, it makes the content relevant for the supervised model to process the data. This can be done by eliminating stopword, noise, and encode consistency. For the prediction of multiclass classification, the transformer model outperformed with an accuracy score of 0.91.(Zerrouki and Benblidia, 2024)

2.3 mBERT Model in Classification

mBert is a modified BERT model, trained with 104 languages and takes advantage of grasping and processes several language data simultaneously.(Panchadara, 2024) Multilanguage content is vital in an electronically connected environment.The classification of offensive and non-offensive comments, the difficulty to detect it in multilingual context is discussed. Taking advantage of the multilingual BERT model, comments are classified in various languages English, Hindi, Telugu, Malayalam, Kannada, Greek, and Russian and achieved an accuracy score of 0.925.(Nandhini et al., 2024) The mBERT model for meme classification outperformed other transformer models and achieved an f1 score of 0.75, 0.95 and 0.71 for Tamil, Malayalam, and Kannada, respectively.(Ghanghor et al., 2021)

3 Proposed Approach

3.1 Problem Overview

Abusive Tamil and Malayalam Text targeting Women on Social Media is addressed as a binary classification problem. The goal is to predict unlabeled instances as Abusive (P=1) or Non-Abusive (P=0). According to the standard probability of supervised learning, the set of output P (0 or 1) is deterministically related to the input N (Text) to output P is denoted by the target function R:N->P. In addition, $N \epsilon n \; and \; P \epsilon p$. (Anthony, 2008) The model is built in such a way that

$$M(N) -> P \tag{1}$$

where P belongs to [0,1].

3.2 About Dataset

The YouTube comments dataset provided by DravidianLangTech, to perform the shared task of promoting Tamil and Malayalam Text targeting Women on Social Media-DravidianLangTech@NAACL 2025(Rajiakodi et al., 2025). The training, validation and evaluation datasets contain data in the low resource languages Tamil and Malayalam. The description

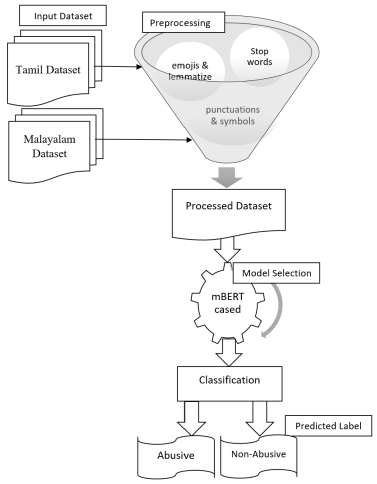


Figure 4: Methodology

Dataset	Abusive	Non-Abusive	Total
Training	1366	1423	2789
Validation	277	320	597
Testing	--	--	598

Table 1: Dataset Description-Tamil

of dataset Tamil and Malayalam is described in Table 1 and Table 2 respectively. The dataset is annotated as abusive or non-abusive according to the context of the comments targeted towards women in social media. The sample dataset of the shared task is shown in Figure 3.
Abusive- Content related to violence, abuse, mistreat, or neglect of women from intimate or dependent relationships or society.
Non-Abusive- Content without abusive context against women.
The distribution of the dataset in languages Tamil and Malayalam is shown in Figure 1 and Figure 2 respectively.

3.3 Preprocessing

Text preprocessing is the transformation of text by cleaning noise and preparing the text for further operation. The various techniques involved are removal of stop words, punctuation, special symbols, and numbers. Preprocessing helps the model to improve the performance of the model(Siino et al., 2024). The content of the dataset, and labels are converted to the lower case (codemix). The preprocessing techniques used are
Removal of noise- includes removal of special symbols, punctuations, emojis.
Tokenization- partitioning the sentences into words.
Lemmatization- Converting the original word to root word.
Removal of stop words- Removal of insignificant words from the instances.
Case conversion- Changing the comments to lowercase.

Dataset	Abusive	Non-Abusive	Total
Training	1530	1402	2932
Validation	303	325	628
Testing	--	--	629

Table 2: Dataset Description-Malayalam

3.4 Evaluation Metrics

The model is validation with the evaluation metrics accuracy, precision, and f1 score. Accuracy(Acc) measures the performance of the model, which is measured by the ratio no of instance to the correctly predicted instances.

$$Acc = \frac{No.of\,flawless\,prediction}{Total\,No.of input\,instance}$$

Prediction(Prec) measures the number of positive predictions that the model considers to be correct. It is calculated by division of the number of true positive(TP) prediction with the sum of true positive and false positive(FP) predictions.

$$Prec = \frac{TP}{TP+FP}$$

Recall(Rec) measures the number of positive prediction obtained by the model is correct. It is calculated by division of the number of true positive(TP) prediction with the sum of true positive, false negative(FN) prediction.

$$Rec = \frac{TP}{TP+FN}$$

True Positive(TP): The predicted output is Yes, and the actual output is Yes as well.
True Negatives(TN): The predicted output and the actual output is also No.
False Positives(FP): The predicted output is Yes, but the actual output is No.
False Negatives(FN): The predicted output is No, but the actual output is Yes.

Language	Accuracy	Precision	f1 score
Tamil	0.77	0.78	0.77
Malayalam	0.72	0.72	0.72

Table 3: mBERT cased-Development Results

3.5 Model selection and implementation

The BERT base multilingual cased is a powerful model, which can be used for 104 languages including Tamil and Malayalam.(Devlin et al., 2018) The articles from Wikipedia are used to train the mBERT model, its performance is based on the quality of the content of the language.(Wu and Dredze, 2020) It is enhanced with a particular training data from a single language, to another language, and enables it to work across different languages.(Nabiilah et al., 2024)

The mBERT model is used for the implementation of the shared task, the model is trained by training the parameters epoch=7, maxlength=256, batch size=16 and AdamW optimizer with the learning rate of 2e-5 and correct bias=True. The result is shown as confusion matrix in Figure 5 and 6. The result is evaluated with the evaluation metrics, accuracy, precision and recall scores of 0.77,0.78, and 0.77 respectively for Tamil. For Malayalam, we obtain the accuracy, precision and recall score of 0.72, 0.72, and 0.72 respectively for development dataset.

4 Results and Discussions

The implementation is performed in Google Colab using the Python programming language with the multilingual BERT model. The model is trained for 7 epochs by tuning the parameters maxlength=256, batch size=16, and AdamW optimizer. The AdamW parameters learning rate and correct bias are set to 2e-5 and True respectively. When the experiment is carried out by setting the correct bias to False for Malayalam language we got biased result for non-abusive label. The different runs are executed with the same parameters for the South Indian languages Tamil and Malayalam. The outperformed results can be viewed in confusion matrix in Figure 5 and 6. The result details of our team, given by the organizers are tabulated in Table 4 and our development result in Table 3. We noticed variation in the model performance related to the number of stop words also. The accuracy of the model dropped when the stopword is increased

Language	Runs	mF1 Score
Tamil	1	0.75
Tamil	2	0.765
Tamil	3	0.757
Malayalam	1	0.674
Malayalam	2	0.677

Table 4: Result Score- DravidianLangTech@NAACL 2025

when working with Tamil language and the reverse for Malayalam language. The usage of lemmatization dropped the accuracy of Tamil language.

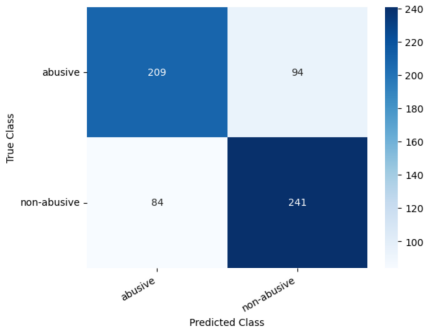


Figure 5: Confusion Matrix-Malayalam

5 Conclusion and Future Work

Usage of abusive text on social networks is a common violation and classification is a challenging task. The dataset shared by DravidianLangTech@NAACL 2025 to classify abusive Tamil and Malayalam Text targeting women in social media is used for the implementation of the model. The pretrained transformer model BERT base multilingual cased was used for the classification in all the runs. The model achieved the mF1 score of 0.765 and 0.677 for Tamil and Malayalam dataset respectively. When working with low resource languages the unavailability of stop words and dictio-

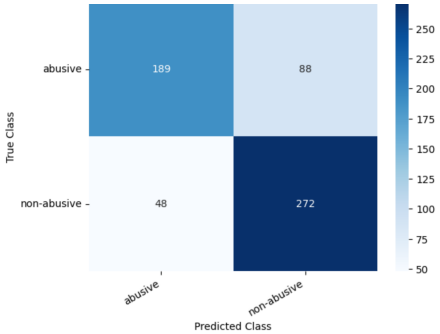


Figure 6: Confusion Matrix-Tamil

nary is another issue. Future research can focus on preprocessing with back translation of sentences, as well as the implementation of vectorization to improve the accuracy of the model.

6 Limitations

Although the implemented model performed well, it has certain limitations. The size of the training dataset is small, which limits the generalization that leads the model to struggle on an unseen dataset. To obtain an accurate result, the ambiguity of words and morphological complexity, which is fairly common in low resource languages such as Tamil and Malayalam, should be addressed.

References

Martin Anthony. 2008. Aspects of discrete mathematics and probability in the theory of machine learning. *Discrete applied mathematics*, 156(6):883–902.

Shubhankar Barman and Mithun Das. 2023. hate-alert@DravidianLangTech: Multimodal abusive language detection and sentiment analysis in Dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 217–224, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Iiitk@ dravidianlangtech-eacl2021: Offensive language identification and meme classification in tamil, malayalam and kannada. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 222–229.

Santoshi Kumari. 2022. Text mining and pre-processing methods for social media data extraction and processing. In *Handbook of research on opinion mining and text analytics on literary works and social media*, pages 22–53. IGI Global.

C Jerin Mahibha, Sampath Kayalvizhi, Durairaj Thenmozhi, and Sundar Arunima. 2021. Offensive language identification using machine learning and deep learning techniques.

Ghinaa Zain Nabiilah, Islam Nur Alam, Eko Setyo Purwanto, and Muhammad Fadlan Hidayat. 2024. Indonesian multilabel classification using indobert embedding and mbert classification. *International Journal of Electrical & Computer Engineering (2088-8708)*, 14(1).

PS Nandhini, R Karunamoorthi, P Mariappan, and S Revathi. 2024. Multilingual offensive language detection in social media content using bert-base-multilingual-cased model. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Kiranmaye Panchadara. 2024. Enhancing named entity recognition in low-resource dravidian languages: A comparative analysis of multilingual learning and transfer learning techniques. *Journal of Artificial intelligence and Machine Learning*, 2(1):1–7.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Pradeep Kumar Roy. 2024. Deep ensemble network for sentiment analysis in bi-lingual low-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–16.

Tiara Intana Sari, Zalfa Natania Ardilla, Nur Hayatin, and Ruhaila Maskat. 2022. Abusive comment identification on indonesian social media data using hybrid deep learning. *IAES International Journal of Artificial Intelligence*, 11(3):895–904.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024. Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, 121:102342.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Khadidja Zerrouki and Nadjia Benblidia. 2024. Multilingual text preprocessing and classification for the detection of extremism and radicalization in social networks.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan,

Yiwei Li, Yifan Zhou, Hanqi Jiang, et al. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.

SSNTrio@DravidianLangTech 2025: LLM Based Techniques for Detection of Abusive Text Targeting Women

Mirnalinee T T
Sri Sivasubramaniya Nadar College of Engineering
MirnalineeTT@ssn.edu.in

Bhuvana J
Sri Sivasubramaniya Nadar College of Engineering
bhuvanaj@ssn.edu.in

Avaneesh Koushik
Sri Sivasubramaniya Nadar College of Engineering
avaneesh2210179@ssn.edu.in

Diya Seshan
Sri Sivasubramaniya Nadar College of Engineering
diya2210208@ssn.edu.in

Rohan R
Sri Sivasubramaniya Nadar College of Engineering
rohan2210124@ssn.edu.in

Abstract

This study focuses on developing a solution for detecting abusive texts on social media against women in Tamil and Malayalam, two low-resource Dravidian languages in South India. As the usage of social media for communication and idea sharing has increased significantly, these platforms are being used to target and victimize women. Hence an automated solution becomes necessary to screen the huge volume of content generated. This work is part of the shared Task on Abusive Tamil and Malayalam Text targeting Women on Social Media DravidianLangTech@NAACL 2025. The approach used to tackle this problem involves utilizing LLM based techniques for classifying abusive text. The Macro Average F1-Score for the Tamil BERT model was 0.76 securing the 11th position, while the Malayalam BERT model for Malayalam obtained a score of 0.30 and secured the 33rd rank. The proposed solution can be extended further to incorporate other regional languages as well based on similar techniques.

1 Introduction

Social media has become an indispensable aspect of our daily lives and has enabled us to remotely communicate with the entire world. It is increasingly involved in delivering important information that shapes people's thoughts and ideologies. However such platforms are easily misused to disseminate abusive content, especially against women. Due to societal biases and gender inequality, women are frequently the target of hateful and demeaning comments that aim to harass or threaten them. Misogyny is the most prevalent form of online hate across all the platforms and about two-thirds of all hateful posts targeted at women were found to be harassment [European Union Agency for Fundamental Rights, 2023]. Such derogatory comments have a severe emotional and psychological impact. Hence it becomes important to ensure that content in such platforms are regulated to avoid biased and harmful content.

With several million videos and comments posted every day, the task of manually classifying abusive comments becomes nearly impossible. To address this challenge, an online content moderation system is essential. Several solutions have been proposed for content moderation using machine learning-based techniques. However, there are limited solutions available for low-resource languages.

The objective of this shared task is to identify abusive comments directed at women in YouTube comments, specifically in Tamil, a language spoken across several parts of Southeast Asia, and Malayalam, which is spoken in certain regions of South India.

2 Related works

Platforms that are faced with the prospect of moderating content face two primary challenges: (1) enforcing policies at scale; (2) ensuring that policies are applied consistently [Schaffner et al., 2024]. A model based approach ensures that policies for moderation of abusive comments against women can be applied at scale.

Tradtional ML based approaches have been widely used for similar use cases. SVM based models are reliable and have achieved good performance in the sentiment classification of harrassments toward women based on Twitter data [Mustapha et al., 2024]. Logistic regression is also commonly used for hatte speech detection in tweets, and the model has achieved high precision, recall, and F1-score for both classes, demonstrating its effectiveness in predicting same [Rathod et al., 2023].

Deep learning approaches that utilize neural networks are becoming increasingly popular alternatives to traditional machine learning techniques as they have the ability to efficiently pick up com-

plicated attributes and context details and a CNN-BiLSTM based approach is proposed by [Vetagiri et al., 2024].

LLM-based approaches, when combined with carefully designed reasoning prompts, can effectively capture the nuanced context of hate speech. By leveraging the extensive knowledge base and contextual understanding of large language models, these methods can accurately identify implicit biases, linguistic subtleties, and evolving patterns of hate speech, significantly outperforming traditional detection techniques [Guo et al., 2024].

3 Dataset Description

The dataset [Priyadharshini et al., 2023] [Priyadharshini et al., 2022] provided for this task [Rajiakodi et al., 2025] comprises comments from various YouTube videos scraped from the internet.

The data was chosen to ensure that they contained controversial and sensitive topics where gender-based abuse is prevalent. It also contains sentences that reflect the colloquial terms that are commonly used in a derogatory manner.

The dataset has two classes, namely abusive and non-abusive comments. The distribution of data points across the two classes is visualized in figure 1 and figure 2. It can be concluded from the figures that both the datasets are balanced.

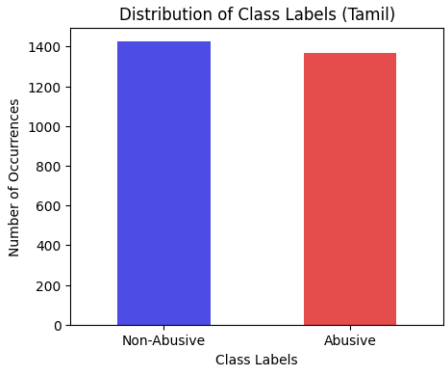


Figure 1: Dataset Description - Distribution of abusive and non-abusive comments in Tamil

4 Methodology

4.1 Data Pre-processing

The corpus consists of text that has undergone an initial preprocessing phase to ensure a cleaner and more structured dataset. In addition to these basic preprocessing steps, further refinement techniques

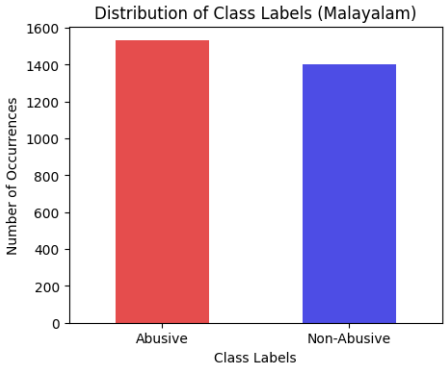


Figure 2: Dataset Description - Distribution of abusive and non-abusive comments in Malayalam

were applied to enhance the text quality. Specifically, special symbols, unnecessary punctuation marks, and non-linguistic characters were systematically removed to eliminate noise while preserving the integrity of the original content. However, stopword removal was intentionally avoided to retain essential linguistic information, ensuring that key contextual and syntactic elements remain intact. This decision was made to prevent the loss of critical words that contribute to sentence structure, meaning, and semantic coherence, particularly in tasks where stopwords play a crucial role in preserving grammatical correctness and contextual nuances.

4.2 Text Tokenization

Tokens are the smallest individual unit of text that represent a meaningful segment of language. Tokenization is done to preprocess textual data by converting raw text into tokenized representations suitable for input into a machine learning or deep learning model. Tokenization was performed on both the datasets using the pre-trained tokenizers of Tamil BERT and Malayalam MBERT [Joshi, 2022].

4.3 Proposed Model

After pre-processing and tokenization, the datasets were randomly split into training and testing data to measure the model's performance. Here 80% of the dataset is considered for training and 20% of the dataset is considered for testing. The models were trained on the dataset and their accuracies were calculated using the test set.

Hyper-parameters that were used for training the LLM models are as follows:

- **Learning Rate** : During training, the learn-

ing rate is a hyperparameter that modifies the model's weights and regulates the step size at each iteration. This was set to 2e-5.

- **Per device Train Batch Size** : The amount of training examples handled on each device (such as a GPU) every training step is determined by the per_device_train_batch_size parameter. This was set to 16.

- **Per device Train Batch Size** : To avoid overfitting, weight decay is a regularization strategy that applies a penalty to the model's loss function according to the magnitude of the weights. This was set to 0.01.

It was observed that most multilingual models exhibited overfitting, as their training errors were significantly lower than their testing errors. Consequently, monolingual BERT models for Tamil and Malayalam were explored, yielding promising results.

To mitigate overfitting, the early stopping technique was used. Early stopping is a widely used technique to prevent overfitting during model training by monitoring the model's performance on a validation dataset. As training advances, the model often increases its performance on both training and validation data sets. However, at some point, the model may begin to memorize the training data rather than learning generalizable patterns. As a result, the validation loss begins to increase while the training loss continues to drop, indicating overfitting.

5 Result and Analysis

Several other models were explored and their training accuracies are highlighted in tables 1 and 2. The monolingual BERT models for Tamil and Malayalam were trained on the dataset of Youtube comments and their results are summarised in table 3. The models for the two languages ranked 11th and 33rd on the leaderboard in their respective subtasks.

The initial model chosen for experimentation was Multilingual-BERT (mBERT), a transformer-based model designed to handle multiple languages. However, its performance was found to be suboptimal when compared to the language-specific BERT models. Although useful for cross-lingual tasks, mBERT's multilingualism seems to limit its capacity to accurately represent the complex language subtleties and contextual connections of Malayalam and Tamil. The accuracies of several other multilingual models explored for Tamil and Malayalam are highlighted in tables 1 and 2.

Model	Training Accuracies
SVM	0.67
Logistic Regression	0.65
ai4bharat indic BERT	0.68
MBERT	0.73
MuRIL	0.74

Table 1: Training Accuracies of the models explored for Tamil

Model	Training Accuracies
SVM	0.63
Logistic Regression	0.62
MBERT	0.69
BERT-base	0.68

Table 2: Training Accuracies of the models explored for Malayalam

Tasks requiring a thorough comprehension of linguistic subtleties can be more effectively performed by language-specific BERT models, especially in languages with intricate morphology, grammatical gender distinctions, and culturally distinctive expressions. Since these models are only trained on one language, they are better able to catch contextual dependencies, idiomatic usage, and complex patterns that are frequently missed by multilingual BERT models.

By focusing solely on one language, monolingual BERT models can leverage a more comprehensive and fine-grained representation of linguistic patterns, leading to improved performance in context-sensitive NLP applications such as hate speech detection, machine translation, and named entity recognition.

Language-Specific BERT Models have the ability to handle context dependent misogynistic and abusive text. BERT's bidirectional attention mechanism helps analyze the surrounding context to infer the true intent of a sentence. The ability of BERT models to perform subword tokenization also helps

Language	Macro F1-Scores
Tamil	0.76
Malayalam	0.30

Table 3: Final results

capture variations in word forms that may indicate that the text is abusive.

Reasons for the poor performance of the Malayalam BERT model could be due to overfitting on training data and due to a the requirement of additional layers on top of BERT as BERT embeddings are rich but may need additional layers to learn task-specific patterns.

6 Conclusion

Language-specific BERT models have proven to be highly effective in addressing tasks that require a deep understanding of linguistic nuances, particularly those related to gender, cultural context, and local expressions. These models are fine-tuned on data from a single language, allowing them to capture language-specific syntactic structures, semantic meanings, and idiomatic expressions that are often deeply intertwined with gender distinctions and subtle social dynamics within the language.

Language-specific models not only excel at gender bias detection but also in tasks that involve understanding cultural connotations, societal norms, and emotional tones in language. This capability is crucial in fields such as sentiment analysis, hate speech detection, and abusive language classification, where contextual meaning plays a significant role.

7 Future Improvements

This work can be expanded in the future by applying similar methodologies and models to other regional languages, enabling a more inclusive and diverse set of language resources. Additionally, leveraging advanced fine-tuning techniques, such as adversarial training, semi-supervised learning, or few-shot learning, can further enhance the model's ability to adapt to different linguistic nuances, handle domain-specific data, and generalize better to unseen or out-of-distribution test data.

The performance of the Malayalam BERT model can be further improved by using techniques like data synthesis or text augmentation which can help improve language coverage. Advanced techniques for tokenization can also be utilized to help the model handle complex linguistic structures efficiently.

8 Limitations

The methodology used in this work can be extended to several low-resource languages, making it adaptable for a wider range of linguistic contexts. However, since language-specific BERT models were utilized, this approach is limited to languages that have a dedicated pre-trained BERT model available. Many low-resource languages lack such models due to insufficient training data, which restricts the scalability of this methodology. In such cases, exploring multilingual models is a suitable option.

9 Ethical Considerations

This research adheres to the ethical guidelines outlined in the ACL Publication Ethics Policy. Efforts were made to ensure that the dataset used in this study respects data privacy, and no personally identifiable information was included. The system is intended to assist humans in responsibly moderating the use of social media and to contribute positively to online safety in real-world applications.

References

European Union Agency for Fundamental Rights. 2023. Online content moderation: Fundamental rights perspectives.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An investigation of large language models for real-world hate speech detection. *Preprint*, arXiv:2401.03346.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Wan Nor Asyikin Wan Mustapha, Norlina Mohd Sabri, Nor Azila Awang Abu Bakar, Nik Marsyahariani Nik Daud, and Azilawati Azizan. 2024. Detection of harassment toward women in twitter during pandemic based on machine learning. *International Journal of Advanced Computer Science and Applications*, 15(3):1035–1041.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth

U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Rutuja G. Rathod, Yashoda Barve, Jatinderkumar R. Saini, and Sourav Rathod. 2023. From data preprocessing to hate speech detection: An interdisciplinary study on women-targeted online abuse. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pages 1–8.

Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. "community guidelines make this the best party on the internet": An in-depth study of online platforms' content moderation policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–16. ACM.

Advaitha Vetagiri, Gyandeep Kalita, Eisha Halder, Chetna Taparia, Partha Pakray, and Riyanka Manna. 2024. Breaking the silence detecting and mitigating gendered abuse in hindi, tamil, and indian english online spaces. *Preprint*, arXiv:2404.02013.

CUET-NLP_MP@DravidianLangTech 2025: A Transformer and LLM-Based Ensemble Approach for Fake News Detection in Dravidian Languages

Md Minhazul Kabir, Md. Mohiuddin
Kawsar Ahmed and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904040, u1904103, u1804017}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Fake news detection is a critical problem in today's digital age, aiming to classify intentionally misleading or fabricated news content. In this study, we present a transformer- and LLM-based ensemble method to address the challenges in fake news detection. We explored various machine learning (ML), deep learning (DL), transformer, and LLM-based approaches on a Malayalam fake news detection dataset. Our findings highlight the difficulties faced by traditional ML and DL methods in accurately detecting fake news, while transformer- and LLM-based ensemble methods demonstrate significant improvements in performance. The ensemble method combining Sarvam-1, Malayalam-BERT, and XLM-R outperformed all other approaches, achieving an F1-score of 89.30% on the given dataset. This accomplishment, which contributed to securing 2^{nd} place in the shared task at DravidianLangTech 2025, underscores the importance of developing effective methods for detecting fake news in Dravidian languages.

1 Introduction

The term fake news describes false or misleading material presented as fact, often shared to deceive and manipulate public opinion. With the rise of social media platforms, the fabrication and rapid spread of fake news have reached unprecedented levels due to the ease of sharing information and the lack of stringent verification processes (Shahi et al., 2021). This phenomenon has significant consequences, including societal division, misinformation, and loss of trust in reliable sources. Fake news can cause serious social, political, and economic issues, including swaying public opinion, upsetting democratic processes, and making it more difficult to handle crises during important occasions like elections or medical crises (Kaliyar et al., 2021). Addressing this issue requires robust approaches to differentiate facts from false narratives, especially in linguistically diverse regions.

Despite significant advancements in detecting fake news across high-resource languages like English, Spanish, and Arabic (Zhou et al., 2023), low-resource languages like Malayalam remain underexplored due to limited datasets and linguistic resources (Thara and Poornachandran, 2022). Malayalam presents unique challenges due to its rich linguistic features, including dialects and idiomatic expressions, making it difficult to process and analyze (Coelho et al., 2023). The rise of code-mixed text, where users mix multiple languages and scripts, further complicates traditional monolingual fake news detection systems (Hegde et al., 2022). This shared task focuses on creating effective approaches for identifying and classifying fake news in Malayalam, emphasizing low-resource and code-mixed scenarios. As a participant in this shared task, our contribution can be summarized as follows:

- Proposed an ensemble approach leveraging two transformer-based models (Malayalam-BERT and XLM-R) and an LLM (sarvam-1) for effective fake news detection in Malayalam.
- Conducted a comprehensive evaluation of various ML models (LR, MNB, SVM, XGBoost), DL models (CNN, LSTM, CNN+BiLSTM), Transformer-based models (Malayalam-BERT, XLM-R, mBERT, DistilBERT), and LLM models (Gemma-2-2b, Llama-3.2-3B, ProjectIndus, sarvam-1) to identify an optimal approach for fake news detection in Malayalam.

2 Related Work

In recent years, a great deal of study has been prompted by the growing ubiquity of fake news across platforms and languages. Coelho et al.

(2023) investigated machine learning models for Dravidian language fake news identification where they trained an ensemble model combining MNB, LR and SVM using TF-IDF of word unigrams that achieved a macro F1-score of 0.83 and third place in the task at DravidianLangTech@RANLP 2023. By utilizing the XLM-RoBERTa base model, renowned for multilingual capabilities, Raja et al. (2023) presented an innovative method that achieved a remarkable macro-averaged F1-score of 87%. Meanwhile, Sujan et al. (2023) employed a multimodal approach by concatenating features from LSTM networks for textual data and VGG16 for image data, achieving a macro F1-score of 0.67. Similarly, Farsi et al. (2024) presented a fine-tuned MuRIL model leveraging parameter tuning that achieved F1-scores of 0.86 and 0.5191 in tasks 1 and 2, respectively, securing 3rd place in task 1 and 1st place in task 2 in DravidianLangTech shared task. In a different research, Devika et al. (2024) curated a dataset specifically for Malayalam fake news detection, achieving an F1-score of 0.3393 with LR trained on LaBSE features while emphasizing the need to address data imbalance. Osama et al. (2024) also contributed to the DravidianLangTech shared task. Their experiments with ML, DL, and transformer-based models revealed that m-BERT achieved the highest macro F1 score of 0.85, securing 4th place in the shared task. Furthermore, Shohan et al. (2024) proposed an intelligent text checkworthiness technique, achieving F1-scores of 75.82% in English (RoBERTa), 52.55% in Arabic (Dehate-BERT), and 58.42% in Dutch (Dutch-BERT). As an instance of notable progress in fake news detection, Kaliyar et al. (2021) presented FakeBERT, which combines BERT with single-layer CNNs, that detected bogus news in English with an impressive 98.90% accuracy. In the context of Arabic fake news detection, Othman et al. (2024) introduced a hybrid model combining 2D-CNN and AraBERT, with F1-scores of 0.6188, 0.7837, and 0.8009 on the ANS, Ara-News, and Covid19Fakes datasets, respectively. Considering the current improvements, this study applies an ensemble method to identify fake news in Dravidian languages.

3 Dataset and Task Description

Task 1 of the shared task competition on "Fake News Detection in Dravidian Languages" (Subramanian et al., 2025, 2024, 2023) focuses on classifying social media posts, specifically YouTube comments into two classes: fake and original in the Malayalam language. The objective is to identify whether a given text is containing misleading or authentic information. The provided dataset (Devika et al., 2024) for this task includes multilingual and codemixed Malayalam data. The given dataset is divided into three sets: train, dev, and test. Each set of data has a nearly equal distribution of fake and original content. Some additional information about the dataset are provided in Table 1.

Set	Class	SC	TW	UW	Avg. Len
Train	original	1658	37229	17472	11
	Fake	1599			
Dev	original	409	8760	5492	7
	Fake	406			
Test	original	512	11266	6587	7
	Fake	507			

Table 1: Dataset distributions, with acronyms SC, TW, UW, and Avg. Len representing sample count, total words, unique words, and average length, respectively.

4 System Overview

This section outlines the methodology for the fake news detection task, which combines traditional machine learning, deep learning, transformer models, and large language models. The diagram of the proposed approach is illustrated in Figure 1. Detailed implementation and source code for this system are accessible on GitHub[1].

In the preprocessing phase, we applied essential text-cleaning techniques such as emoji removal, HTML tag removal, duplicate sample removal, etc. For feature extraction, we used different methods tailored to the type of model being employed. TF-IDF (Takenobu, 1994) and the CountVectorizer method was used extensively for ML models to represent the importance of terms based on their frequency across the documents. For DL models, we employed GloVe embeddings (Pennington et al., 2014), which capture the semantic relationships and contextual meaning of words in the text.

4.1 Machine Learning Approaches

Several machine learning models such as SVM, XGBoost, MNB, LR, and ensemble methods were explored for the task. LR was used with the regularization parameter set to 0.1 and a maximum of 50000 iterations. SVM was implemented using a linear kernel with a regularization parameter

[1] https://github.com/R1FA7/Fake-News-Detection-Malayalam

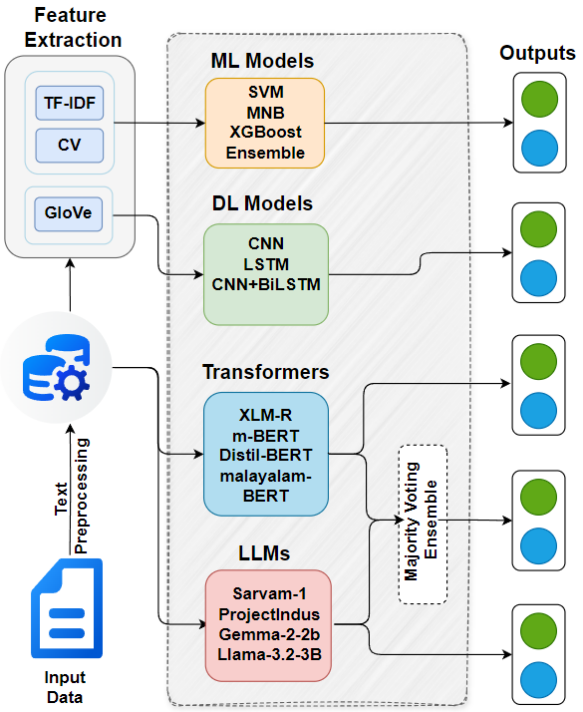


Figure 1: Schematic process for fake news detection

C set to 0.8 and a gamma value of 1. MNB was tested with an alpha value of 1.0, which controls the smoothing of probabilities. Also, XGBoost was implemented with eval metric mlogloss. Finally, an ensemble model combining those was employed, using a majority voting scheme for classification.

4.2 Deep Learning Approaches

For deep learning, we explored CNN, LSTM, BiLSTM architectures to capture the intricate patterns in the text data. A simple CNN was used, consisting of one convolutional layer with 128 filters, followed by GlobalMaxPooling for dimensionality reduction and a dense layer for classification, optimized with Adam and binary cross-entropy loss. LSTM networks were also used to capture sequential dependencies in the text, with a batch size of 64 and a single LSTM layer comprising 200 units. A hybrid CNN and LSTM model was explored combining CNN's feature extraction and BiLSTM's (Schuster and Paliwal, 1997) bidirectional context to improve performance.

4.3 Transformer Based Approaches

We leveraged several pre-trained transformer-based models to address the fake news classification task from Hugging Face (Wolf, 2020). These models included m-BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), Distil-BERT (Sanh et al., 2020), and Malayalam-BERT (Joshi, 2023) model. As part of finetune, we concatenated the train and dev set data first. This work randomly took 15% for validation and the rest for training to maximize the data available for training the data-hungry transformer models. Reserving 15% data for validation allowed for hyperparameter tuning and model evaluation without significantly reducing the training data. The text data was tokenized and padded to a maximum sequence length of 90 tokens. The models were trained using a learning rate of $3e^{-5}$, a batch size of 16, and a total of 10 epochs. The performance was optimized based on the macro F1 score.

4.4 LLM Based Approaches

We also explored state-of-the-art Large Language Models (LLMs) such as Gemma-2-2b (Team, 2024), ProjectIndus (Malhotra et al., 2024), sarvam-1 [2] and Llama-3.2-3B (Liu et al., 2024) to further enhance the performance. To fine-tune these models we combined the train and dev sets, randomly kept 20% for validation, and used the rest for training, along with various optimization strategies. First, we leveraged 4-bit quantization to reduce the model's memory footprint without affecting its performance. For model adaptation, we employed Parameter Efficient Fine-Tuning (PEFT), particularly focusing on LoRA (Low-Rank Adaptation). LoRA helps fine-tune large models with fewer parameters, making the process more efficient by updating only key attention layers and freezing other layers. This approach allows the model to adapt to the task without extensive retraining. The LoRA setup used a rank of 4, with a scaling factor (alpha) of 16 and a dropout of 0.15 ensuring that the model's performance improved without overfitting. For the training process, we set the learning rate to $1e^{-4}$ utilizing the AdamW optimizer with a weight decay of 0.01 and batch size 16 in 5 epochs. Additionally, a learning rate scheduler with a reduction factor of 0.5 and patience of 2 epochs was applied to maintain steady improvement.

4.5 Ensemble Approaches

In the ensemble approaches, we employed a majority voting scheme to combine the predictions of multiple models in order to enhance the overall performance. The majority voting technique works by aggregating the predictions from different models

[2] https://huggingface.co/sarvamai/sarvam-1

and selecting the class that appears most frequently as the final decision. This approach benefits from the strengths of different models, increasing robustness and accuracy. For our ensemble model, we experimented with different combinations of models from various categories, such as Transformer-based models and Large Language Models. Specifically, we combined models like Malayalam-BERT, XLM-R, mBERT and Distil-BERT from the transformer category with sarvam-1 from the LLM category. These models were selected based on their strong individual performance, ensuring that each contributed unique strengths to the ensemble.

5 Results and Analysis

The results in Table 2 reveal notable differences in the performance of various ML, DL, Transformer, and LLM models. In this study, we used the G-mean score instead of precision and recall to ensure balanced performance across both classes, avoiding bias even in balanced datasets.

Among the ML models, the Ensemble methods with countVectorizer features achieved the highest macro F1-score of 77.11%, outperforming other ML models. In DL, CNN with GloVe embeddings performed the best, achieving an F1-score of 61.00%, followed by CNN+BiLSTM at 61.00%. However, these scores were significantly lower than the top-performing ML models. Among the transformer-based models, Malayalam-BERT outperformed all other models with an F1-score of 88.32%, surpassing XLM-R (86.36%) and mBERT (85.27%) by a notable margin. Distil-BERT scored an F1-score of 84.00%. In the LLMs category, sarvam-1 achieved the highest F1-score of 83.90%, outperforming Google's Gemma-2-2B (82.63%) and Meta's Llama-3.2-3B (83.27%) by a small margin. Finally, our proposed model, an ensemble of sarvam-1, Malayalam-BERT, and XLM-R, achieved the highest performance with an F1-score of 89.30%, which is 1.2% higher than the next best-performing transformer (Malayalam-BERT), and 12.26% higher than the best ML model (Ensemble+CV).

ML and DL models delivered lower scores compared to Transformer and LLM models. Specifically, The score of DL was significantly lower than ML which can be due to their struggle to generalize well on smaller or limited datasets, leading to overfitting on the training data. Malayalam-BERT is a BERT model trained on publicly available Malayalam monolingual datasets, which excelled due to its specialization in the Malayalam language. LLM didn't perform like the transformer-based models possibly due to their less specialized nature for this specific task or less fine-tuning due to resource limitations. However, in our proposed method, We chose the top two transformer models (Malayalam-BERT and XLM-RoBERTa) and the best LLM model (sarvam-1) for a majority voting ensemble, as they showed the best individual performance. This combination excels by combining the complementary strengths of two transformers and one LLM model, delivering superior performance compared to other model ensembles.

ML Models			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
SVM+TF-IDF	74.94	75.04	75.00
XGBoost+CV	74.50	73.25	74.00
MNB+CV	75.34	75.62	76.00
LR+CV	76.13	76.31	76.00
Ensemble+ CV	**77.46**	**77.11**	**77.00**
DL Models			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
LSTM(GloVe)	56.92	60.00	63.00
CNN+ BiLSTM(GloVe)	57.71	61.00	64.00
CNN(GloVe)	**58.15**	**61.00**	**64.00**
Transformers			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
XLM-R	86.49	86.36	86.00
mBERT	84.50	85.27	85.00
Distil-BERT	83.64	83.61	83.63
Malayalam-BERT	**88.50**	**88.32**	**88.00**
LLMs			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
Gemma-2-2b	82.45	82.63	82.51
Llama-3.2-3B	84.00	83.27	84.00
ProjectIndus	59.97	59.86	60.00
sarvam-1	**83.98**	**83.90**	**84.00**
Ensemble			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
(mBERT + XLM-R + Malayalam-BERT)	88.12	88.10	88.10
(Distil-BERT +XLM-R + Malayalam-BERT)	88.15	88.13	88.15
(sarvam-1 + XLM-R + Malayalam-BERT) (Proposed)	**89.48**	**89.30**	**89.40**

Table 2: Performance of the different methods on the test set

5.1 Error Analysis

A comprehensive quantitative and qualitative error analysis is conducted to provide detailed insights into the proposed model's performance.

Quantitative Analysis

Figure 2 presents a confusion matrix that classifies text in the test set as either fake or original. The figure indicates that out of 1019 test samples, 910 were correctly identified and 109 were incorrect predictions. Figure 3 shows that the model was trained on Fake sentences with an average length of 14.51, while Original sentences had an average

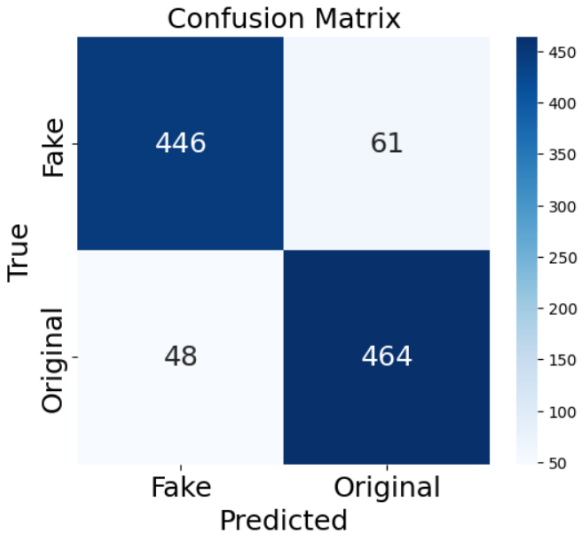


Figure 2: Confusion matrix of Ensemble model

length of 8.46. However, when the model encounters Fake sentences in the test set with an average length of 7.77, it tends to misclassify them. This discrepancy in sentence length suggests that the model may be struggling with shorter Fake sentences, leading to ambiguity and incorrect predictions.

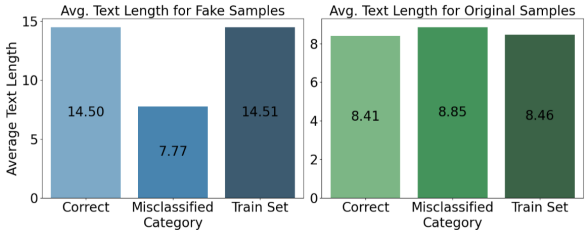


Figure 3: Classified and misclassified average length of classes

Qualitative Analysis

Figure 4 presents some predicted outputs of the developed model. In the first and third texts, the model successfully predicted the class of the text. On the other hand, it failed to do so in the second and fourth texts. The proposed model, which is fine-tuned in this work, is primarily trained on long length fake data compared to fake data in test data. This could be one of the reasons for this model's failure in some samples.

6 Conclusion

This study investigated the performance of several LLM, transformer-based, DL, and ML models on the Malayalam fake news detection dataset. The results demonstrate that the the ensemble model combining sarvamai/sarvam-1, Malayalam-BERT, and XLM-RoBERTa reached the highest F1-score of 89.30%. This finding shows the effectiveness of the transformer and LLM-based ensemble approach, in successfully tackling the challenge of fake news detection. Fine-tuning on shorter fake news samples, and exploring advanced preprocessing methods for code-mixed data could further boost the performance in the future.

Text	Actual	Predicted
വിഷമം വരബോൾ ഞാൻ ഓടി വരും (I will run when trouble comes)	Original	Original
നശിപ്പിച്ചു കളയണം ഈ കമ്മ്യുണിസ്റ്റ് രാജ്യത്തെ (This communist country must be destroyed)	Fake	Original
Corona ye jesus oddikkum (Corona is Jesus' Holy Spirit)	Fake	Fake
വാക്സിൻ എന്തിനാണ്? (Why the vaccine?)	Original	Fake

Figure 4: Sample predictions made by the proposed Ensemble model with actual and predicted label.

Limitations

Despite the effectiveness of our ensemble method, several limitations remain. One major concern is the reliance on transformer-based models and large language models, which may not generalize well to languages beyond Malayalam. Additionally, the model's performance was affected by the limited training data, particularly with shorter fake news samples, leading to misclassifications. Future work can address these challenges by expanding the training dataset with more diverse and representative samples, improving fine-tuning strategies, and exploring cross-lingual transfer learning to enhance generalization across different languages. Integrating more advanced data augmentation techniques and leveraging multimodal approaches may also contribute to improved robustness and accuracy in fake news detection.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Sharal Coelho, Asha Hegde, G Kavya, and Hosahalli Lakshmaiah Shashirekha. 2023. Mucs@ dravidianlangtech2023: Malayalam fake news detection using machine learning approach. In *Proceedings of*

the Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 288–292.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Salman Farsi, Asrarul Eusha, Ariful Islam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. Cuet_binary_hackers@ dravidianlangtech eacl2024: Fake news detection in malayalam language leveraging fine-tuned muril bert. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 173–179.

Asha Hegde, Shubhanker Banerjee, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Hosahalli Shashirekha, John Philip McCrae, et al. 2022. Overview of the shared task on machine translation in dravidian languages. In *Proceedings of the second workshop on speech and language technologies for Dravidian languages*, pages 271–278.

Raviraj Joshi. 2023. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *Preprint*, arXiv:2211.11418.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2024. Spinquant: Llm quantization with learned rotations. *Preprint*, arXiv:2405.16406.

Nikhil Malhotra, Nilesh Brahme, Satish Mishra, and Vinay Sharma. 2024. Project indus: A foundational model for indian languages. *Tech Mahindra Makers Lab*.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Cuet_nlp_goodfellows@ dravidianlangtech eacl2024: A transformer-based approach for detecting fake news in dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192.

Nermin Abdelhakim Othman, Doaa S Elzanfaly, and Mostafa Mahmoud M Elhawary. 2024. Arabic fake news detection using deep learning. *IEEE Access*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. volume 14, pages 1532–1543.

Eduri Raja, Badal Soni, and Sami Kumar Borgohain. 2023. nlpt malayalm@ dravidianlangtech: Fake news detection in malayalam using optimized xlm-roberta model. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 186–191.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Mike Schuster and Kuldip Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681.

Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of covid-19 misinformation on twitter. *Online social networks and media*, 22:100104.

Symom Hossain Shohan, Md Sajjad Hossain, Ashraful Islam Paran, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Semanticcuetsync at checkthat! 2024: Pre-trained transformer-based approach to detect check-worthy tweets.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Adhish S Sujan, Aleena Benny, VS Anoop, et al. 2023. Malfake: A multimodal fake news identification for malayalam using recurrent neural networks and vgg-16. *arXiv preprint arXiv:2310.18263*.

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Gemma Team. 2024. Gemma.

S Thara and Prabaharan Poornachandran. 2022. Social media text analytics of malayalam–english code-mixed using deep learning. *Journal of big Data*, 9(1):45.

Thomas Wolf. 2020. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Lina Zhou, Jie Tao, and Dongsong Zhang. 2023. Does fake news in different languages tell the same story? an analysis of multi-level thematic and emotional characteristics of news about covid-19. *Information Systems Frontiers*, 25(2):493–512.

CUET-NLP_Big_O@DravidianLangTech 2025: A Multimodal Fusion-based Approach for Identifying Misogyny Memes

Md. Refaj Hossan, Nazmus Sakib, Md. Alam Miah
Jawad Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904007, u1904086, u1904102, u1704039}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Memes have become one of the main mediums for expressing ideas, humor, and opinions through visual-textual content on social media. The same medium has been used to propagate harmful ideologies, such as misogyny, that undermine gender equality and perpetuate harmful stereotypes. Identifying misogynistic memes is particularly challenging in low-resource languages (LRLs), such as Tamil and Malayalam, due to the scarcity of annotated datasets and sophisticated tools. Therefore, DravidianLangTech@NAACL 2025 launched a Shared Task on Misogyny Meme Detection to identify misogyny memes. For this task, this work exploited an extensive array of models, including machine learning (LR, RF, SVM, and XGBoost), and deep learning (CNN, BiLSTM+CNN, CNN+GRU, and LSTM) are explored to extract textual features, while CNN, BiLSTM + CNN, ResNet50, and DenseNet121 are utilized for visual features. Furthermore, we have explored feature-level and decision-level fusion techniques with several model combinations like MuRIL with ResNet50, MuRIL with BiLSTM+CNN, T5+MuRIL with ResNet50, and mBERT with ResNet50. The evaluation results demonstrated that BERT + ResNet50 performed best, obtaining an F1 score of 0.81716 (Tamil) and were ranked 2nd in the task. The early fusion of MuRIL+ResNet50 showed the highest F1 score of 0.82531 and received a 9th rank in Malayalam.

1 Introduction

The unprecedented proliferation of social media has brought about an exponential increase in meme-based communication, where image and text combine in powerful messages of influence in public opinion (Singh et al., 2024). Memes are primarily vehicles for humor and social commentary (Ponnusamy et al., 2024); however, in multilingual contexts, they have increasingly been used as a conduit for misogynistic content (Suryawanshi et al., 2020b; a P K et al., 2020). Misogynistic memes are digital seeds of negativity disguised as humor that perpetuate harmful stereotypes and normalize disrespect toward women (H et al., 2024; Singh et al., 2024). Hence, determining whether shared content on social media is misogynistic or not is necessary.

Although much recent work has explored the analysis of emotions conveyed in memes (Mishra et al., 2023), the identification of offensive and hate content in memes (Hermida and Santos, 2023; Rizwan et al., 2024), focused on identifying misogynistic content in memes in high-resource languages such as English (Farinango Cuervo and Parde, 2022). The challenge remains in LRLs, such as Tamil and Malayalam, due to a lack of resources (Magueresse et al., 2020), linguistic complexity, and cultural quirks (Kumari et al., 2023). To address these challenges, the DravidianLangTech@NAACL 2025 Shared Task on Misogyny Meme Detection[1] focused on Tamil and Malayalam languages (Ponnusamy et al., 2024; Chakravarthi et al., 2025), two widely spoken Dravidian languages with distinct linguistic characteristics. The challenge aims to develop a robust approach to identify misogynistic content in memes in these languages. It requires processing visual and textual information and grasping their combined meaning in cultural and linguistic contexts. In this paper, we take an integrated approach to this challenge, leveraging the power of advanced models in both visual and textual modalities. Hence, the contributions of the work are as follows:

- Developed a multimodal architecture that effectively combines visual and textual features for misogynistic content detection in Tamil and Malayalam languages.

[1] https://codalab.lisn.upsaclay.fr/competitions/20856

- Investigated various ML, DL, and transformer-based models with different fusion techniques to identify misogynistic memes while evaluating performance metrics and conducting error analysis to determine the best strategy for detecting misogynistic content in both languages.

2 Related Work

While the meme culture is going strong, there has been significant research into detecting trolling, hostility, offensive, and abusive language from social media data, with several studies conducted by researchers (Suryawanshi et al., 2020b,a; Kumari et al., 2023; H et al., 2024) in recent years. Many studies have focused solely on textual features to detect harmful content (Sreelakshmi et al., 2020; Baruah et al., 2020). However, many researchers have investigated memes' textual and visual features to classify trolls, offenses, and aggression. For instance, Suryawanshi et al. (2020a) introduced the MultiOFF[2] meme dataset, containing 743 memes for detecting offensive content and proposed a stacked LSTM with VGG16 approach for multi-modal analysis, which outperformed other models using a single feature, achieving an F1-score of 0.50. Another study by Sultan et al. (2024) introduced MemesViTa, a multi-modal fusion model combining Vision Transformer (ViT) and DeBERTa, which achieved 94.29% accuracy and 95.82% F1 score, surpassing both visual and textual models in troll meme detection. However, in the context of misogyny meme identification, several works on misogyny meme identification focus on a linguistic perspective (Anzovino et al., 2018), proposing a corpus of misogynistic tweets and exploring machine learning models for detection. Butt et al. (2021) tackled sexism detection in multilingual social media text, achieving an F1 score of 0.78 for sexism identification and 0.49 for categorization using data augmentation. A data augmentation approach using song lyrics was introduced to improve misogyny detection, outperforming conventional transfer learning techniques on English and Spanish datasets (Calderón-Suarez et al., 2023).

Several studies have revealed multi-modal approaches for misogyny content detection. Ponnusamy et al. (2024) developed the MDMD (Misogyny Detection Meme Dataset) to analyze misogyny, gender bias, and stereotypes in Tamil and Malayalam-speaking communities through memes. Rizzi et al. (2023) evaluated four uni-modal and three multi-modal approaches for detecting misogynistic memes, introducing a bias estimation technique and Bayesian Optimization to improve accuracy by 61.43%. Another work done by Singh et al. (2024) achieved a 0.73 F1-score on 5,054 Hindi-English memes using BiT+MuRIL, outperforming unimodal models. H et al. (2024) used MNB for text and ResNet50 for images, achieving F1-scores of 0.69 (Tamil) and 0.82 (Malayalam) in LT–EDI 2024[3] shared task. Although hateful and offensive memes are well-studied, misogynistic meme detection in Tamil and Malayalam remains unexplored. This work improves previous efforts by integrating visual and textual modalities for better performance.

3 Task and Dataset Description

This study aims to detect misogynistic content in memes using textual and visual features, framing it as a binary classification task in Tamil and Malayalam languages. The given dataset (Ponnusamy et al., 2024; Chakravarthi et al., 2024) contains a total of 1776 data points in the Tamil language and 1000 data points in the Malayalam language, combining train, validation, and test set. Table 1 shows the class-wise distribution of samples for the Tamil dataset, highlighting an imbalance with more *Non-misogyny* samples (851 train, 210 valid, 267 test) compared to *Misogyny* samples (285 train, 74 valid, 89 test).

Classes	**Train**	**Valid**	**Test**	$\mathbf{W}_T$	$\mathbf{UW}_T$
Non-misogyny	851	210	267	26770	15194
Misogyny	285	74	89	9243	6749
Total	1136	284	356	36013	21943

Table 1: Class-wise distribution of train, validation, and test set for the Tamil language, where W_T and UW_T denote total words and total unique words in three datasets.

Table 2 presents the class-wise distribution for the Malayalam dataset, showing an imbalance with more *Non-misogyny samples* (381 train, 97 valid, 122 test) compared to *Misogyny* samples (259 train, 63 valid, 78 test). This imbalance, coupled with the smaller size and vocabulary of the *Misogyny* class,

[2] https://shorturl.at/DEyxx

[3] https://codalab.lisn.upsaclay.fr/competitions/16097

poses challenges for model performance.

Classes	Train	Valid	Test	W_T	UW_T
Non-misogyny	381	97	122	11004	7574
Misogyny	259	63	78	6398	4378
Total	640	160	200	17402	11952

Table 2: Class-wise distribution of train, validation, and test set for the Malayalam language, where W_T and UW_T denote total words and total unique words in three datasets.

The implementation details of the tasks will be found in the GitHub repository[4].

4 Methodology

Several ML, DL, and transformer-based models were explored to develop a framework for Misogyny meme detection (Figure 1).

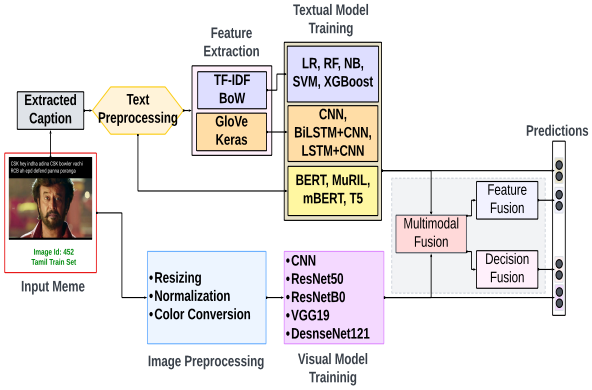


Figure 1: Schematic process of Misogyny meme detection.

4.1 Data Preprocessing

Text preprocessing included language-specific tokenization, such as MuRIL tokenizer for Malayalam and Tamil, with a maximum sequence length of 128 tokens, handling special characters, preserving language-specific Unicode characters, and eliminating extra whitespace. Image preprocessing involved the most common transformations, such as scaling images to 224x224 pixels, converting to RGB format, and normalizing using ImageNet statistics *(mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225])*. The preprocessing pipeline maintains different configurations for the training and inference phases.

[4] https://github.com/RJ-Hossan/MMD-NAACL-2025

4.2 Feature Extraction

TF-IDF and Bag of Words (BoW) were used to extract textual features for ML models. In TF-IDF, up to 5000 features were extracted, and stop words were removed. DL models used 100-dimensional pre-trained GloVe embeddings. An embedding matrix was built to map vocabulary words to vectors. For out-of-vocabulary terms, zero vectors were used. Text tokenization and vocabulary creation were performed using a *CountVectorizer* with 7000 features, while contextual embeddings were reduced using a fully connected layer. For images, pre-trained models extracted visual features aligned with the pre-processing pipeline.

4.3 Baselines

Several unimodal (visual or textual) and multimodal (visual and textual) models were explored and fused using early and late fusion approaches, with the necessary hyperparameter tuning, to perform the tasks.

4.3.1 Unimodal Baselines

Various ML and DL models were utilized to develop the unimodal approach. For textual features, traditional models such as LR, SVM, RF, and Gradient Boosting and deep learning models such as BiLSTM, CNN, and BiLSTM+CNN were employed, using GloVe and Keras-based embedding. CNN-based architectures, such as DenseNet-121, EfficientNet-B0, ResNet-50, and VGG19, were implemented for visual features.

Table 3 outlines the hyperparameters for the LR, RF, SVM, and XGBoost models trained on textual features. LR uses *max_iter=1000* and *random_state=27* whereas Random Forest sets *n_estimators=100* and *random_state=27*. The SVM used a specified kernel. XGBoost adopted *eval_metric=mlogloss*. Each model has different parameter configurations.

Parameter	LR	RF	SVM	XGBoost
max_iter	1000	-	-	-
random_state	27	27	15	15
n_estimators	-	100	-	-
kernel	-	-	linear	-
probability	-	-	True	-
use_label_encoder	-	-	-	False
eval_metric	-	-	-	mlogloss

Table 3: Parameters used for ML models (textual features only).

Table 4 presents the tuned hyperparameters for deep learning (DL) models using textual features, including BiLSTM+CNN, Text-CNN, LSTM+CNN, and BiLSTM. It outlines parameters such as *embedding size*, *max sequence length, hidden dimension, filter sizes, number of layers, dropout rate, learning rate*, and *optimizer (Adam)*. The loss function varies between *CrossEntropyLoss* and *BinaryCrossEntropyLoss*.

Hyperparameter	BiLSTM+CNN	Text-CNN	LSTM+CNN	BiLSTM
embedding_size	100 (GloVe vectors)	100 (GloVe vectors)	100 (GloVe vectors)	100 (GloVe vectors)
max_sequence_length	100	100	100	100
hidden_dimension	256	-	256	256
number_of_filters	-	128	128	-
filter_sizes	-	[3, 4, 5]	[3]	-
number_of_layers	2	-	-	2
batch_size	32	32	32	32
learning_rate	0.001 (Adam)	0.001 (Adam)	0.001 (Adam)	0.001 (Adam)
epochs	45	45	45	45
dropout_rate	-	0.5	-	-
loss_function	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss
optimizer	Adam	Adam	Adam	Adam

Table 4: Tuned hyperparameters for DL models (textual features only).

Table 5 outlines the tuned hyperparameters for DL models using visual features, including CNN, VGG19, EfficientNetB0, and DenseNet121. It specifies *input size, preprocessing techniques, learning rate, loss function, batch size, activation function (ReLU, Sigmoid),* and *fine-tuning strategy* (whether the layers are frozen or not).

Hyperparameter	CNN	VGG19	EfficientNetB0	DenseNet121
Input Size	(224, 224, 3)	(224, 224, 3)	(224, 224, 3)	(224, 224, 3)
Base Model	-	VGG19	EfficientNetB0	DenseNet121
Optimizer	Adam	Adam	Adam	Adam
Learning Rate	0.001	0.001	0.001	0.001
Loss Function	Binary Crossentropy	Binary Crossentropy	Binary Crossentropy	Binary Crossentropy
Epochs	20	20	20	20
Batch Size	32	32	32	32
Activation Function	ReLU, Sigmoid	ReLU, Sigmoid	ReLU, Sigmoid	ReLU, Sigmoid
Dropout Rate	0.5	0.5	0.5	0.5
Fine-Tuning	No (frozen layers)	No (frozen layers)	No (frozen layers)	No (frozen layers)

Table 5: Tuned hyperparameters for DL models (visual features only).

4.3.2 Multimodal Baselines

The multimodal baseline models combined textual and visual features using early fusion (EF) and late fusion (LF) techniques. Models like T5+MuRIL+ResNet-50, BiLSTM+CNN+MuRIL, MuRIL+ResNet-50, and BERT+ResNet-50 utilized pre-trained models for feature extraction. For instance, the text is tokenized in Tamil using BERT with padding and truncation to a fixed length of 128 tokens and processed through a fully connected layer to reduce dimensionality to 256. The images were resized to 224x224 pixels and processed through ResNet-50, with the final fully connected layer replaced to output 256 features. Textual and visual features were concatenated into a single vector of size 512, which was passed through a classifier. The training used the AdamW optimizer with a learning rate of 2e-5, a batch size of 16, and a learning rate scheduler that reduced the rate when validation loss plateaus. Despite differences in dataset size, class distribution, and language complexity between Tamil and Malayalam, the tuned hyperparameters for both tasks were kept similar and are presented in Table 6. Moreover, similar models, i.e., BiLSTM+CNN+MuRIL, MuRIL+ResNet-50, and T5+MuRIL+ResNet-50, were also employed with late fusion, where fusion happened only at the decision level after both modalities had been independently processed.

Hyperparameter	BERT+ResNet50 (Tamil)	MuRIL+ResNet50 (Malayalam)
Learning Rate	2e-5	2e-5
Batch Size	16	16
Number of Epochs	45	45
Max Sequence Length	128	128
Optimizer	AdamW	AdamW
Dropout Rate	0.3	0.3
Image Model	ResNet50	ResNet50
Text Model	BERT (base-uncased)	MuRIL (base)
Scheduler	ReduceLROnPlateau	ReduceLROnPlateau

Table 6: Tuned hyperparameters used in multimodal fusion for Tamil and Malayalam languages.

4.4 System Requirements

Most models, specifically fusion models, were trained on a dual GPU setup (NVIDIA Tesla T4x2), utilizing parallel processing for textual and visual features. The BERT+ResNet50 model used 7-8 GB of GPU memory, while MuRIL+ResNet50 required approximately 8-10 GB of GPU memory. Training for 45 epochs for both approaches took 120-150 minutes, depending on the size of the data set and the calculation of the class weight.

5 Result Analysis

Table 7 provides a comparative analysis of the performance of different approaches used to detect misogynistic memes in Tamil and Malayalam datasets. In textual-only methodologies, conventional models such as SVM and Gradient Boosting performed well, with F1 scores of 0.6507 and 0.6848 (Tamil) and 0.6693 and 0.7058 (Malayalam). As for deep learning models, BiLSTM (Glove) had the lowest F1 scores (0.4826 Tamil, 0.5187 Malayalam), while BiLSTM+CNN (Glove) performed best for Tamil (0.6703) and LSTM+CNN (Glove) for Malayalam (0.6448).

Regarding visual-only models, EfficientNet-B0 performed better in both languages, achieving an F1 score of 0.6546 for Tamil and 0.7640 for Malayalam. The VGG19 model had shown excellent performance, especially for the Malayalam language,

Approaches	Classifiers	Tamil				Malayalam			
		P	R	F1	G	P	R	F1	G
Textual Only	LR	0.7871	0.5618	0.5496	0.6650	0.7128	0.6803	0.6858	0.6964
	SVM	0.7027	0.6348	0.6507	0.6679	0.6727	0.6673	0.6693	0.6700
	RF	0.7720	0.5805	0.5807	0.6694	0.6895	0.6963	0.6911	0.6929
	Gradient Boosting	0.7197	0.6635	0.6848	0.6910	0.7058	0.7058	0.7058	0.7058
	BiLSTM (Glove)	0.4750	0.5000	0.4826	0.4873	0.5750	0.5167	0.5187	0.5451
	CNN (Glove)	0.6347	0.6367	0.6357	0.6357	0.5922	0.5632	0.5514	0.5775
	LSTM+CNN (Glove)	0.6420	0.6592	0.6480	0.6505	0.6884	0.6836	0.6448	0.6860
	BiLSTM+CNN (Glove)	0.6774	0.6648	0.6703	0.6711	0.6426	0.6318	0.6305	0.6372
Visual Only	CNN	0.6146	0.5787	0.5844	0.5964	0.6619	0.6568	0.6587	0.6593
	DenseNet-121	0.8072	0.5787	0.5786	0.6835	0.7120	0.6693	0.6739	0.6903
	EfficientNet-B0	0.8332	0.6330	0.6546	0.7262	0.7616	0.7722	0.7640	0.7669
	ResNet-50	0.6782	0.5843	0.5899	0.6295	0.8134	0.8058	0.8007	0.8096
	VGG19	0.7448	0.6854	0.7044	0.7145	0.8134	0.8058	0.8215	0.8096
Multi-modal Fusion (EF)	T5+MuRIL+ResNet-50	0.8170	0.8146	0.8158	0.8158	-	-	-	-
	BiLSTM+CNN+MuRIL	0.8365	0.7865	0.8065	0.8111	0.8270	0.8255	0.8262	0.8262
	MuRIL+ResNet-50	0.8281	0.7981	0.8013	0.8130	**0.8451**	**0.8157**	**0.8253**	**0.8303**
	BERT+ResNet-50	**0.8160**	**0.8184**	**0.8172**	**0.8172**	-	-	-	-
Multi-modal Fusion (LF)	T5+MuRIL+ResNet-50	0.8315	0.7940	0.8099	0.8125	-	-	-	-
	BiLSTM+CNN+MuRIL	0.8160	0.7884	0.8005	0.8021	0.8374	0.8226	0.8284	0.8299
	MuRIL+ResNet-50	0.8178	0.8031	0.8017	0.8104	0.8044	0.7911	0.7962	0.7977
	mBERT+EfficientNet-B0	-	-	-	-	0.7844	0.7857	0.7851	0.7850

Table 7: Result comparison on test data, where EF, LF, P, R, F1, and G denote early fusion, late fusion, precision, recall, F1-score, and geometric mean score of precision and recall, respectively.

achieving an F1 score of 0.8215. Concerning multimodal fusion, EF models like BERT+ResNet-50 performed better than others in Tamil with an F1 score of 0.8172, which helped us rank 2nd in this task. During our observation after the competition, we noted that the BiLSTM+CNN+MuRIL model outperformed the MuRIL+ResNet-50 model in Malayalam with a higher F1 score of 0.8262 but had a lower G score[5] of 0.8262 compared to 0.8303. In addition to this, the LF models performed well, as BiLSTM+CNN+MuRIL reached an F1-score of 0.8284 for the Malayalam language. Appendix A demonstrates a comprehensive error analysis of the employed models.

6 Conclusion

This paper demonstrated a shared task solution to detect misogynistic memes in Tamil and Malayalam that exploited textual and visual characteristics. The results showed that the multimodal approach BERT+ResNet-50 with early fusion achieved the highest F1 score of 0.8172 in Tamil. However, MuRIL with ResNet50 outperformed all models and obtained the highest F1 score (0.8253) through early fusion in Malayalam. Although the results are promising, the current approach has room for further improvement. Advanced preprocessing techniques, such as data augmentation, would enrich the dataset and improve the generalization of the model. Future work aims to explore vision-based transformer models and advanced multimodal techniques (i.e., CLIP) for enhanced performance.

7 Limitations

The study on misogynistic content identification in multimodal memes has several drawbacks, influenced by following factors:

- Fine-tuned DL and transformer models may fail when meme contexts differ from training data.
- Since the dataset used for training the models was imbalanced and no advanced augmentation was employed, it could have led to biased predictions in another set of memes.
- Although multimodal fusion approaches showed strong results, the complexity of combining multiple models and managing text-image interactions may have caused computational inefficiencies and overfitting, limiting scalability.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

[5]The G score of precision and recall is the square root of the product of precision and recall.

References

Abdul Rasheed a P K, Carmel Jose, and Anju Michael. 2020. Social media and meme culture: A study on the impact of internet memes in reference with 'kudathai murder case'.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. *Automatic Identification and Classification of Misogynistic Language on Twitter*, pages 57–64.

Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Aggression identification in English, Hindi and Bangla text using BERT, RoBERTa and SVM. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 76–82, Marseille, France. European Language Resources Association (ELRA).

Sabur Butt, Noman Ashraf, Alexander Gelbukh, and Grigori Sidorov. 2021. Sexism identification using bert and data augmentation–exist2021.

Ricardo Calderón-Suarez, Rosa M. Ortega-Mendoza, Manuel Montes-Y-Gómez, Carina Toxqui-Quitl, and Marco A. Márquez-Vera. 2023. Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases. *IEEE Access*, 11:13179–13190.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Charic Farinango Cuervo and Natalie Parde. 2022. Exploring contrastive learning for multimodal detection of misogynistic memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 785–792, Seattle, United States. Association for Computational Linguistics.

Shaun H, Samyuktaa Sivakumar, Rohan R, Nikilesh Jayaguptha, and Durairaj Thenmozhi. 2024. Quartet@LT-EDI 2024: A SVM-ResNet50 approach for multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 221–226, St. Julian's, Malta. Association for Computational Linguistics.

Paulo Cezar de Q. Hermida and Eulanda M. dos Santos. 2023. Detecting hate speech in memes: a review. *Artificial Intelligence Review*, 56(11):12833–12851.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2023. Emoffmeme: identifying offensive memes by leveraging underlying emotions. *Multimedia Tools and Applications*, 82:1–36.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *ArXiv*, abs/2006.07264.

Shreyash Mishra, Suryavardan S, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Naquee Rizwan, Paramananda Bhaskar, Mithun Das, Swadhin Satyaprakash Majhi, Punyajoy Saha, and Animesh Mukherjee. 2024. Zero shot vlms for hate meme detection: Are we there yet? *Preprint*, arXiv:2402.12198.

Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing Management*, 60(5):103474.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

K. Sreelakshmi, Premjith B., and Soman Kp. 2020. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744.

Tipu Sultan, Mohammad Abu Tareq Rony, Mohammad Shariful Islam, Saad Aldosary, and Walid El-Shafai. 2024. Memesvita: A novel multimodal fusion technique for troll memes identification. *IEEE Access*, 12:177811–177828.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020b. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

A Error Analysis

We conducted quantitative and qualitative error analyses to gain a deeper understanding of the performance of the best-performed model.

Quantitative Analysis: The best-performing models were used for a quantitative error analysis, utilizing confusion matrices for Tamil and Malayalam to identify misogynistic memes. Figure A.1 demonstrated that the proposed BERT+ResNet50 model using late fusion revealed strong overall performance with an accuracy of 86.2%.

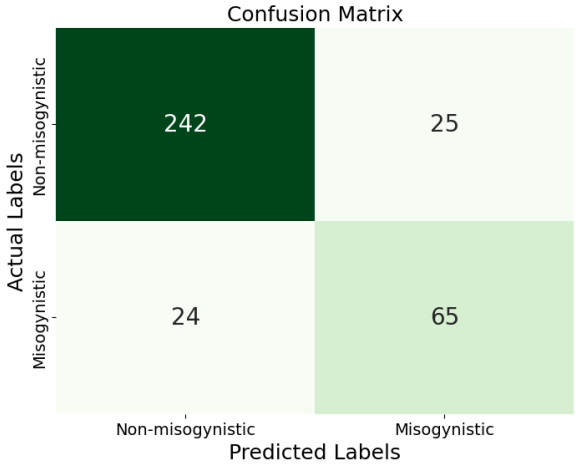


Figure A.1: Confusion matrix of the proposed approach (BERT+ResNet50 by early fusion) for Tamil language.

The model correctly identifies 242 *non-misogynistic* memes and 65 *misogynistic* memes while misclassifying 25 *non-misogynistic* memes as *misogynistic* and missing 24 *misogynistic* memes. The misclassifications arise from visually complex memes with overlapping misogynistic and non-misogynistic elements and subtle text based on sarcasm or cultural context. We have found that the model struggles with imbalanced data, leading to the misclassification of most non-misogynistic memes as misogynistic due to the subtlety in textual complexity.

Figure A.2 shows the confusion matrix to identify misogyny memes for the Malayalam language, using the fusion of MuRIL (for text) and ResNet50 (for images). It shows that 113 *non-misogynistic* memes are correctly identified, while 55 *misogynistic* memes are accurately identified as *misogynistic* from 78 misogynistic memes. However, 9 *non-misogynistic* memes are incorrectly classified as *misogynistic*, and 23 *misogynistic* memes are missed. These outcomes indicate that the model performs relatively well but still has some drawbacks due to poor handling of data imbalance and the dynamic nature of the contextual meanings of memes.

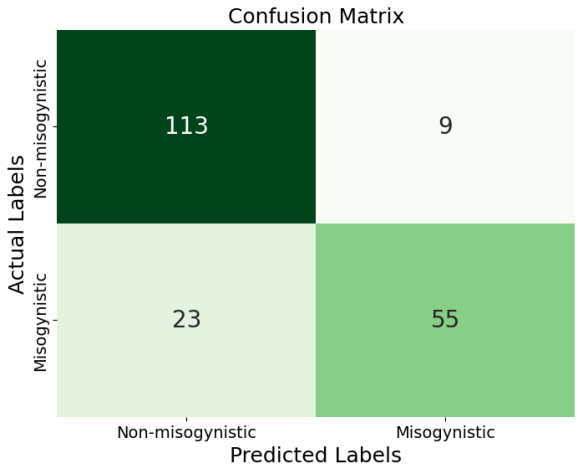


Figure A.2: Confusion matrix of the proposed approach (MuRIL+ResNet50 by early fusion) for Malayalam language.

Qualitative Analysis: Figure A.3 and A.4 highlight the best-performed model's predicted outputs for sample inputs in identifying misogynistic memes for both Tamil and Malayalam datasets. In Figure A.3, the proposed model (BERT+ResNet50) accurately predicted samples 2 and 3 but incorrectly predicted samples 1 and 4 in Tamil, indicating some prediction inconsistencies. Similarly, Figure A.4 illustrates the model's performance in Malayalam, where it correctly identified samples 2 and 3 but incorrectly predicted samples 1 and 4. These errors may be due to class imbalance, as the *misogyny* class in the Malayalam dataset contains only 78 instances, which likely impacts the model's generalization capability.

Image Id: 1064.jpg
True Label: Misogyny (1)
Predicted Label: Non-misogyny (0)

Sample 1

Image Id: 1258.jpg
True Label: Misogyny (1)
Predicted Label: Misogyny (1)

Sample 2

Image Id: 1396.jpg
True Label: Misogyny (1)
Predicted Label: Misogyny (1)

Sample 3

Image Id: 1563.jpg
True Label: Non-misogyny (0)
Predicted Label: Misogyny (1)

Sample 4

Image Id: 545.jpg
True Label: Misogyny (1)
Predicted Label: Non-misogyny (0)

Sample 1

Image Id: 488.jpg
True Label: Misogyny (1)
Predicted Label: Misogyny (1)

Sample 2

Image Id: 431.jpg
True Label: Misogyny (1)
Predicted Label: Misogyny (1)

Sample 3

Image Id: 954.jpg
True Label: Non-misogyny (0)
Predicted Label: Misogyny (1)

Sample 4

Figure A.3: Few sample predictions by the BERT+ResNet50 for the Tamil language.

Figure A.4: Some predicted outputs by the MuRIL+ResNet50 for the Malayalam language.

LexiLogic@DravidianLangTech 2025: Detecting Misogynistic Memes and Abusive Tamil and Malayalam Text Targeting Women on Social Media

Niranjan Kumar M[1], Pranav Gupta[1], Billodal Roy[1], Souvik Bhattacharyya[1]
[1]Lowe's
Correspondence: {niranjan.k.m, pranav.gupta, billodal.roy,souvik.bhattacharyya}@lowes.com

Abstract

Social media platforms have become a significant medium for communication and expression, but they are also plagued by misogynistic content targeting women. This study focuses on detecting misogyny in memes and abusive textual content in Tamil and Malayalam languages, which are underrepresented in natural language processing research. Leveraging advanced machine learning and deep learning techniques, we developed a system capable of identifying misogynistic memes and abusive text. By addressing cultural and linguistic nuances, our approach enhances detection accuracy and contributes to safer online spaces for women. This work also serves as a foundation for expanding misogyny detection to other low-resource languages, fostering inclusivity and combating online abuse effectively.

This paper presents our work on detecting misogynistic memes and abusive Tamil and Malayalam text targeting women on social media platforms. Leveraging the pretrained models l3cube-pune/tamil-bert and l3cube-pune/malayalam-bert, we explored various data cleaning and augmentation strategies to enhance detection performance. The models were fine-tuned on curated datasets and evaluated using accuracy, F1-score, precision, and recall. The results demonstrated significant improvements with our cleaning and augmentation techniques, yielding robust performance in detecting nuanced and culturally-specific abusive content.

Our model achieved macro F1 scores of 77.83/78.24 on L3Cube-Bert-Tamil and 78.16/77.01 on L3Cube-Bert-Malayalam, ranking 3rd and 4th on the leaderboard. For the misogyny task, we obtained 83.58/82.94 on L3Cube-Bert-Malayalam and 73.16/73.8 on L3Cube-Bert-Tamil, placing 9th in both. These results highlight our model's effectiveness in low-resource language classification.

1 Introduction

The rise of social media has enabled open communication but has also fueled an increase in toxic and abusive content, with misogyny targeting women becoming a critical concern. Harmful memes and abusive text perpetuate gender-based discrimination, impact mental health, and reinforce societal stereotypes. This issue is particularly challenging in underrepresented languages like Tamil and Malayalam, where cultural nuances and linguistic complexity hinder effective detection. Existing detection tools, designed primarily for high-resource languages, often fail to address the needs of Tamil and Malayalam speakers. Identifying such content requires specialized models that account for low-resource language challenges and cultural context.

This study focuses on detecting misogynistic memes and abusive Tamil and Malayalam text targeting women. By leveraging advanced machine learning models and incorporating cultural insights, we aim to expand NLP capabilities to underserved languages and contribute to safer, more inclusive digital platforms. We present two distinct tasks: (1) detecting misogynistic memes and (2) detecting abusive Tamil and Malayalam text specifically targeting women. Both tasks rely on a combination of linguistic nuance and contextual understanding, which are critical for effective detection.

As part of this effort, we align our work with the NAACL 2025 Dravidian language competitions (Chakravarthi et al., 2025; Rajiakodi et al., 2025), which emphasize the development of NLP solutions for low-resource Dravidian languages. These competitions serve as a platform to advance research in underrepresented languages, fostering collaboration between linguists, computer scientists, and AI researchers. By participating in these shared tasks, we aim to benchmark our models against state-of-the-art approaches and contribute to the broader initiative of improving abusive con-

tent detection in Tamil and Malayalam. Our work not only enhances the technological landscape for Dravidian languages but also promotes responsible AI applications in social media moderation. [1].

2 Related Work

The detection of misogynistic and abusive content has gained importance in NLP due to increasing toxicity on digital platforms. While significant progress has been made for high-resource languages like English, low-resource languages such as Tamil and Malayalam remain underexplored. This survey highlights key efforts in abuse detection, misogyny identification, and multimodal meme analysis, with a focus on Tamil and Malayalam.

Abuse and Hate Speech in Multilingual Contexts: Initial studies, such as (Waseem and Hovy, 2016) and (Wulczyn et al., 2017), advanced abuse detection in English, while (Bhattacharya et al., 2019) addressed multilingual challenges with Hindi-English datasets. Tamil and Malayalam gained attention through (Zhao and Tao, 2021), who introduced annotated datasets for Dravidian languages.

Misogyny Detection: Early research focused on textual misogyny detection (Öhman et al., 2018) and later expanded to multimodal content. For Tamil and Malayalam, leveraged pre-trained models to incorporate cultural nuances in misogyny detection.

Abusive Language Detection in Tamil and Malayalam: Recent work by (Chakravarthi et al., 2021) used transformer models like BERT for gender-specific abuse detection, tackling challenges such as dialect diversity and linguistic richness.

Multimodal Meme Classification: (Kiela et al., 2021) introduced multimodal models combining text and image embeddings, later adapted by for regional languages. Tamil and Malayalam memes, however, lack annotated datasets, limiting progress.

Low-Resource NLP Challenges: Scarcity of labeled data, dialectal complexity, and code-mixing are persistent issues(Hande et al., 2022). Approaches like transfer learning and pre-trained models, such as l3cube-pune/tamil-bert and l3cube-pune/malayalam-bert, have shown potential in overcoming these limitations (Litake et al., 2022).

Conclusion: Despite progress, Tamil and Malayalam remain underrepresented in misogyny and abuse detection research. Leveraging pre-trained models and multimodal techniques can bridge this gap, enabling safer and more inclusive online spaces.(Vaswani et al., 2023), (Devlin et al., 2019)

3 Methodology of processing

3.1 Dataset of IndicBERT

IndicXNLI(Aggarwal et al., 2022) is a benchmark dataset designed to evaluate Natural Language Inference (NLI)(Chen et al., 2018) for 11 major Indian languages, including Hindi, Tamil, Malayalam, Telugu, Kannada, Bengali, Gujarati, Punjabi, Marathi, Oriya, and Assamese. Extending the XNLI dataset(Conneau et al., 2018), it provides premise-hypothesis pairs translated into these languages to capture the linguistic and cultural diversity of the Indian subcontinent. Each instance in the dataset is labeled as entailment, contradiction, or neutral, enabling cross-lingual and multilingual evaluation. IndicXNLI supports the fine-tuning and evaluation of multilingual models such as mBERT, XLM-RoBERTa, and IndicBERT, focusing on challenges like morphological richness, dialectal variations, and code-mixed text common in Indian languages. It serves as a vital resource for assessing NLI performance in low-resource languages while identifying the limitations of existing models in handling complex linguistic structures and cultural nuances. This dataset bridges the gap in NLI research for Indian languages and informs the development of robust and inclusive NLP systems.

3.2 Tamil-BERT and Malayalam-BERT

The L3Cube-Tamil-BERT and L3Cube-Malayalam-BERT models, pre-trained on large-scale corpora specific to Tamil and Malayalam, have demonstrated strong performance on various NLP tasks. When fine-tuned on tasks such as sentiment analysis, named entity recognition, and abusive language detection, these models achieved high accuracy, surpassing many general multilingual models. For instance, L3Cube-Tamil-BERT has shown accuracies of up to 92% on sentiment classification tasks, while L3Cube-Malayalam-BERT achieved around 90% accuracy in the same domain. These models excel in understanding the unique syntactic and semantic structures of Tamil and Malayalam, improving performance on downstream tasks compared to

[1]The code for this paper is available at this GitHub repository

traditional models like mBERT or XLM-R.

In abusive language detection, these models have been shown to yield accuracy rates of around $85 - 88\%$, significantly outperforming other language-specific models that were not fine-tuned for Tamil and Malayalam. By capturing intricate language patterns and cultural context, L3Cube models provide an essential foundation for building advanced NLP systems in low-resource Indian languages, contributing to more effective applications in social media content moderation and sentiment analysis in regional languages.

3.3 Binary Classification

We conducted experiments on two significant tasks from the Dravidian Language Technology Workshop: (1) Offensive Language Identification in Dravidian Languages (Codalab Competition: [20701] and (2) Meme Classification for Tamil (Codalab Competition: [20856]. Both tasks aim to address pressing issues in regional language processing, including offensive language and meme-based toxicity detection, using Tamil and Malayalam as representative low-resource languages.

3.3.1 Offensive Language Identification (Competition 20701)

For this task, we used L3Cube-Tamil-BERT and L3Cube-Malayalam-BERT, fine-tuning them on the provided annotated datasets. The datasets consisted of social media text annotated as offensive or non-offensive. Our pre-processing involved tokenization, cleaning unwanted symbols, and normalizing code-mixed data. The models were fine-tuned with a learning rate of 2e-5 for five epochs using a cross-entropy loss function. On the test set, L3Cube-Tamil-BERT achieved macro F1 score of 78.24, and L3Cube-Malayalam-BERT achieved an F1 score of 70.01, demonstrating their ability to understand nuanced linguistic patterns in offensive content. placing it among the top-performing submissions for this task(Litake et al., 2022)..

3.3.2 Meme Classification for Tamil and Malayalam (Competition 20856)

In this task, we focused on the classification of Tamil and Malayalam memes into categories such as offensive, humorous, or neutral. We utilized L3Cube-Tamil-BERT and L3Cube-Malayalam-BERT for the textual data extracted from memes, and incorporated data augmentation techniques to improve class balance. Pre-trained Tamil-BERT and Malayalam-BERT embeddings provided contextual understanding, crucial for recognizing nuanced meanings within the textual content of memes. Our model achieved macro-F1 score of 68.707 on Tamil-BERT and 80.364 on Malayalam-BERT on test data set(Litake et al., 2022)..

Model	**Train set macro F1**	**Test set macro F1**
XLM-Bert-Tamil	72.01	73.17
Indic-Tamil	73.62	75.04
L3Cube-Bert-Tamil	**77.83**	**78.24**
XLM-Bert-Malayalam	73.52	74.01
Indic-Malayalam	73.92	74.73
L3Cube-Bert-Malayalam	**78.16**	**77.01**

Table 1: Offensive Language Identification Results on Tamil and Malayalam

Observations and Insights The results demonstrate the strength of the L3Cube models in handling low-resource Indian languages. By leveraging domain-specific embeddings, the models captured the linguistic and cultural nuances of Tamil and Malayalam, outperforming baseline multilingual models like mBERT and XLM-R. These experiments highlight the potential of language-specific pre-trained models in advancing NLP tasks for low-resource languages, contributing to safer and more inclusive digital ecosystems.

4 Results

Among the other models, IndicBERT demonstrated competitive performance, with Indic-Tamil and Indic-Malayalam achieving macro F1 scores of 75.04 and 74.73, respectively. XLM-BERT, despite being a widely used multilingual model, exhibited slightly lower performance in comparison. This suggests that models pre-trained specifically on Tamil and Malayalam data, such as L3Cube BERT and IndicBERT, have a notable advantage in handling linguistic intricacies for offensive language identification. The overall results emphasize the effectiveness of domain-specific BERT models in low-resource languages like Tamil and Malayalam. L3Cube BERT, with its targeted pretraining, outperformed general multilingual models, making it a strong candidate for applications in sentiment analysis, offensive language detection, and other NLP

Model	Train set macro F1	Test set macro F1
XLM-Bert-Malayalam	78.5	78.67
Indic-Malayalam	79.83	79.98
L3Cube-Bert-Malayalam	**83.58**	**82.94**
XLM-Bert-Tamil	67.06	69.43
Indic-Tamil	69.01	70.3
L3Cube-Bert-Tamil	**73.16**	**73.8**

Table 2: Meme Classification Results on Tamil and Malayalam

tasks in these languages. Future research could further explore hybrid approaches or fine-tuning strategies to enhance these models' performance further.

5 Conclusion

Our study highlights the significance of leveraging specialized pre-trained models such as l3cube-tamil-bert and l3cube-malayalam-bert for misogyny and abusive content detection in Tamil and Malayalam. Compared to general-purpose multilingual models like XLM-BERT and IndicBERT, the L3Cube models demonstrated superior performance, as evidenced by the results summarized in the comparison tables. These improvements underscore the importance of adopting language-specific embeddings that capture the linguistic and cultural nuances of Tamil and Malayalam.

Additionally, applying data augmentation techniques, such as synonym replacement, back-translation, and contextual data augmentation, contributed significantly to enhancing the models' performance. These methods enriched the training datasets, enabling the models to better generalize to varied and complex scenarios. By combining fine-tuned L3Cube models with robust data augmentation strategies, we achieved improved accuracy and contextual understanding, paving the way for more effective detection of misogynistic and abusive content. This work emphasizes the need for tailored approaches to address the challenges of low-resource languages, fostering safer and more inclusive digital platforms.

6 Limitations

Offensive content against specific groups such as women is a major concern on social media. While our models help in detecting inappropriate content, they rely on static datasets, which might no longer hold valid due to changing trends. Therefore, we need approaches such as active learning and continual learning for ensuring that such offensive content detectors stay up to date and have a balanced representation from new and existing social media platforms. Larger unsupervised and supervised datasets can also improve the performance metrics of such systems, especially in lower resource languages such as Tamil and Malayalam. Another important issue is that of bias- such models might inadvertently discriminate against certain social media users. Moreover, bad actors might exploit the limitations of these models to circumvent NLP-based offensive content detectors, and discover creative ways to post undesirable content.

References

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. Indicxnli: Evaluating multilingual inference for indian languages. *Preprint*, arXiv:2204.08776.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Fire 2019 aila track: Artificial intelligence for legal assistance. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 4–6, New York, NY, USA. Association for Computing Machinery.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Adeep Hande, Siddhanth U Hegde, Sangeetha S, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. The best of both worlds: Dual channel language modeling for hope speech detection in low-resourced Kannada. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 127–135, Dublin, Ireland. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *Preprint*, arXiv:2005.04790.

Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. L3Cube-MahaNER: A Marathi named entity recognition dataset and BERT models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 29–34, Marseille, France. European Language Resources Association.

Emily Öhman, Kaisla Kajava, Jörg Tiedemann, and Timo Honkela. 2018. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Brussels, Belgium. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yingjia Zhao and Xin Tao. 2021. ZYJ@LT-EDI-EACL2021:XLM-RoBERTa-based model with attention for hope speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 118–121, Kyiv. Association for Computational Linguistics.

CUET-NLP_Big_O@DravidianLangTech 2025: A BERT-based Approach to Detect Fake News from Malayalam Social Media Texts

Nazmus Sakib*, **Md. Refaj Hossan***, **Alamgir Hossain**
Jawad Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904086, u1904007}@student.cuet.ac.bd, alamgir.hossain.cs@gmail.com
u1704039@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

The rapid growth of digital platforms and social media has significantly contributed to spreading fake news, posing serious societal challenges. While extensive research has been conducted on detecting fake news in high-resource languages (HRLs) such as English, relatively little attention has been given to low-resource languages (LRLs) like Malayalam due to insufficient data and computational tools. To address this challenge, the DravidianLangTech 2025 workshop organized a shared task on fake news detection in Dravidian languages. The task was divided into two sub-tasks, and our team participated in Task 1, which focused on classifying social media texts as original or fake. We explored a range of machine learning (ML) techniques, including Logistic Regression (LR), Multinomial Naïve Bayes (MNB), and Support Vector Machines (SVM), as well as deep learning (DL) models such as CNN, BiLSTM, and a hybrid CNN+BiLSTM. Additionally, this work examined several transformer-based models, including m-BERT, Indic-BERT, XLM-Roberta, and MuRIL-BERT, to exploit the task. Our team achieved 6th place in Task 1, with MuRIL-BERT delivering the best performance, achieving an F1 score of 0.874.

1 Introduction

In the digital era, social media platforms such as Facebook, Twitter, and Instagram have transformed how people share and consume information. These platforms let users stay updated on current events, express opinions, and participate in real-time global discussions. However, alongside these benefits, the rise of social media has also facilitated the proliferation of false or misleading information, commonly referred to as *fake news* (Subramanian et al., 2023). This phenomenon has become a critical concern due to its far-reaching consequences on public perception, societal trust, and decision-making processes. Fake news is content purposely created to misinform or deceive its audience, often impersonating reputable news sources (Subramanian et al., 2024). The rapid spread of fake news on social media exploits anonymity and platform reach, often outpacing factual content. The effects are severe, resulting in societal divisiveness, a loss of trust in credible news sources, and increased worry among individuals. Furthermore, bogus news can sway political decisions, harm reputations, and exacerbate existing societal divides (Farsi et al., 2024). Although significant progress has been achieved in detecting fake news in resourceful languages like English, less attention has been put towards low-resource languages, such as Malayalam, despite its speakers' rising digital footprint (Sharif et al., 2021). The lack of sufficient annotated datasets and the linguistic complexity of Malayalam pose unique challenges to building reliable fake news detection systems for this language. A shared task was organized under DravidianLangTech@NAACL 2025[1] to address this pressing issue, focusing on classifying social media texts into two categories: *Original* and *Fake* (Devika et al., 2024). As there is little research on Malayalam, we faced various difficulties like linguistic variations, dialect, and semantic identity (Coelho et al., 2023). The primary objective of this research is to design an efficient system capable of accurately classifying Malayalam news samples as fake or original, thus contributing to combating misinformation in low-resource languages. To achieve these objectives, our contributions to the task are as follows:

- Developed a transformer-based framework to detect fake news within the Malayalam dataset.

*Authors contributed equally to this work.

[1] https://sites.google.com/view/dravidianlangtech-2025/home

- Investigated various ML, DL, and transformer-based models, evaluating their performance across metrics to identify the most effective model for detecting fake news in Malayalam. Presented an in-depth error analysis to refine the findings further.

2 Related Work

The proliferation of fake news on platforms like Facebook and Twitter often leads to misinformation and incorrect judgments. This growing concern has paved the way for research leveraging various ML and DL models in this domain (Sharif et al., 2021). While significant efforts have been made to address this issue, limited attention has been given to LRLs such as Malayalam. Different ML approaches have been devoted to a Malayalam dataset by Coelho et al. (2023) for fake news detection. They achieved the highest F1-score of 0.8310 using an ensemble of models (MNB+LR+SVM). In another study, M. San Ahmed (2021) developed a Kurdish dataset and applied ML models like LR, SVM, and Naive Bayes, with SVM achieving the highest accuracy of 88.17%. Additionally, Kumar and Singh (2022) employed ML models on a Hindi dataset containing 2,100 news articles to detect fake news, with Long Short-Term Memory (LSTM) achieving the highest F1-score of 0.89.

A recent study Krešňáková et al. (2019) utilized a fake news dataset from a competition and applied a CNN model, achieving an impressive F1-score of 0.97. Similarly, Kong et al. (2020) explored various neural network models on an English dataset, obtaining an accuracy of 90%. In another study, Kumar et al. (2020) developed a dataset by collecting data from Twitter, where a CNN+BiLSTM model with an attention mechanism achieved an accuracy of 88%. Additionally, Hiramath and Deshpande (2019) employed a dataset comprising news articles and found that a Deep Neural Network (DNN) achieved the highest accuracy of 91%. Several BERT variants, such as XLNet and ALBERT, outperformed deep learning approaches on a COVID-19 dataset (Gundapu and Mamidi, 2021). Schütz et al. (2021) utilized the *FakeNewsNet* dataset (Shu et al., 2020) and applied multiple transformer models, ultimately achieving the best F1 score of 0.84 with RoBERTa. Qazi et al. (2020) compared hybrid CNN models with transformer-based models, finding a slight improvement in F1 score to 0.47 with the transformer models. MuRiL-BERT also performed well on a Telugu dataset, achieving an F1 score of 0.87 (Hariharan et al., 2024). In another study, a comprehensive dataset for fake news detection in Bangla, a low-resource language, was developed, with LLM achieving the best F1 score of 0.89 (Shibu et al., 2025). A key limitation of past studies is their focus on HRLs, which results in biased models that may not transfer well to LRLs like Malayalam. In this context, we have presented a transformer-based framework tailored to handle Malayalam's unique linguistic and cultural aspects, improving detection accuracy for this underrepresented language.

3 Task and Dataset Description

The shared task[2] organizers provided a benchmark dataset for fake news detection in Malayalam (Subramanian et al., 2025). The dataset contains two classes: *Fake* and *Original*. The *Fake* class includes targeted texts, posts, or comments containing misinformation or falsified content, often created to mislead readers for political, commercial, or malicious purposes. The goal is to identify such content, which is especially common on social media during critical events. On the other hand, the class *Original* includes accurate, truthful posts providing reliable, verified information. The dataset contains 3,257 training samples, 815 development samples, and 1,019 test samples. Table 1 illustrated the class-wise distribution of the dataset.

Classes	**Train**	**Dev**	**Test**	**W$_T$**	**UW$_T$**
Original	1658	409	512	14031	8100
Fake	1599	406	507	23198	13100
Total	3257	815	1019	37229	19465

Table 1: Class-wise distribution of the dataset, where W$_T$ and UW$_T$ denote total words in three datasets and total unique words in train data.

The task's goal is to distinguish genuine news from fake news effectively. Figures A.1 and A.2 in Appedix A exhibit the word cloud distribution of classes.

4 Methodology

Several ML, DL, and transformer-based models are implemented and investigated to address the tasks. Figure 1 shows an outline of the methodology.

[2] https://codalab.lisn.upsaclay.fr/competitions/20698

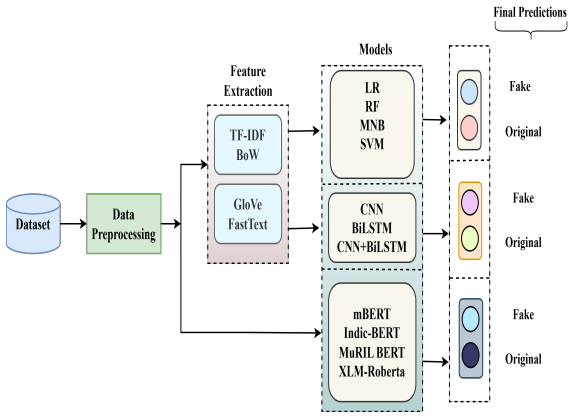


Figure 1: Schematic process of detecting fake news in Malayalam.

4.1 Data Preprocessing

Several preprocessing steps were applied to enhance the dataset's interpretability for the employed models. These steps included cleaning the text and removing unnecessary punctuation, emojis, and hyperlinks that could introduce noise into the data. Additionally, the MuRIL tokenizer was utilized to preprocess the text effectively. We used the MuRIL tokenizer with a maximum sequence length of 128 tokens.

4.2 Feature Extraction

For ML models, we utilized Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) representations with n-grams (unigrams and bigrams). We limited the vocabulary to the top 10,000 terms to balance interpretability and computational efficiency. We leveraged pretrained word embeddings such as GloVe and FastText for DL models. Specifically, we used GloVe embeddings with a dimensionality of 120, which effectively captured word semantics and contextual relationships, and FastText embeddings trained on subword information to handle out-of-vocabulary words. We also employed MuRIL, a transformer-based model that tokenized the Malayalam text with a maximum token length of 128 and provided contextualized embeddings with 768 dimensions. These diverse embedding techniques ensured a robust text representation, enabling the models to accurately identify patterns and distinguish between fake and original news.

4.3 Classifiers

Four ML, six DL, and four transformer-based baselines are explored for fake news detection tasks.

4.3.1 ML Baselines

LR, SVM, RF, and MNB are utilized for the downstream task. The LIBLINEAR (Fan et al., 2008) solver function is used for ML models with GridSearchCV[3] to obtain better results. These traditional machine learning models serve as strong baselines to compare against transformer-based approaches, providing insight into the effectiveness of different learning paradigms. By leveraging GridSearchCV, we systematically tune hyperparameters to optimize each model's performance, ensuring a fair evaluation. This comparative analysis helps assess whether deep learning methods significantly outperform classical techniques in identifying misinformation.

4.3.2 DL Baselines

We employed CNN and the hybrid CNN+BiLSTM model for fake news detection, leveraging their ability to capture spatial and sequential patterns in textual data. The CNN model was designed to extract local features from the text using convolutional filters. Table 2 shows the fine-tuned hyperparameters for the deep learning-based models for the task.

Parameter	Value
Embedding Dimensions	128
Sequence Length	100
CNN Filters	64 filters of size 5
BiLSTM Units	64
Epochs	130
Batch Size	32
Optimizer	Adam
Learning Rate	1e-4

Table 2: Hyperparameter settings for CNN + BiLSTM model.

In contrast, the CNN+BiLSTM hybrid model combined the strengths of CNN's feature extraction with BiLSTM's ability to capture long-term dependencies and context. In our CNN+BiLSTM model, we configured a vocabulary size of 10,000, a sequence length of 100, and an embedding dimension of 128 for tokenization and embedding. The CNN branch was equipped with 64 filters of size 5 for local pattern extraction, and the BiLSTM branch had 64 units to capture bidirectional sequential relationships. The models were trained using

[3] https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html

sparse categorical cross-entropy as the loss function and the Adam optimizer with a learning rate of 1e-4. To address class imbalance, we computed class weights, ensuring fair model performance across categories.

4.3.3 Transformer-based Models

Transformer-based models were employed for fake news detection due to their ability to efficiently process large-scale contextual information, making them well-suited for multilingual tasks (Devlin et al., 2019). Several transformer models, including Indic-BERT (Dabre et al., 2022), mBERT (Pires et al., 2019), XLM-RoBERTa (Zhao and Tao, 2021), and MuRIL-BERT (Khanuja et al., 2021), were explored to evaluate their performance across diverse linguistic settings. Each model was fine-tuned for the classification task, with hyperparameters optimized to enhance performance. The MuRIL-BERT model demonstrated the best results, achieving an F1 score of 0.874. Table 3 presents the fine-tuned hyperparameters for the MuRIL-BERT model.

Parameter	Value
Batch Size	16
Epochs	9
Weight Decay	0.003
Learning Rate	2e-4

Table 3: Hyperparameter configuration for the transformer-based approach (MuRIL-BERT).

The model was trained using a learning rate of 2e-4, a weight decay of 0.003, and for 9 epochs. We trained on 9 epochs, as training for too many epochs (e.g., 10 or 15) led to overfitting, in which the model learns patterns too specific to the training data and loses generalization to unseen data. The optimal results highlight the effectiveness of MuRIL-BERT in handling the complex nature of fake news detection in multilingual datasets.

Additional implementation details can be accessed via the GitHub repository[4].

4.4 System Requirements

The model was trained on a dual GPU setup (NVIDIA Tesla T4x2), utilizing parallel processing for convolutional, BiLSTM, and transformer layers. The CNN+BiLSTM model required 5–8 GB of GPU memory and took approximately 60 minutes to complete training over 130 epochs. In contrast, the MuRIL-BERT model, which required 20 GB of GPU memory, completed training in just 20 minutes for 9 epochs. The training duration varied depending on the dataset size and the computation of class weights for handling class imbalances.

[4] https://github.com/Arghya-n/DravidianLangTech-FakeNews-2025

5 Result Analysis

Table 4 compares the performance of various classifiers for fake news detection, highlighting the precision (P), recall (R), F1 score, and G score.

Classifiers	Fake News Detection			
	P	R	F1	G-Score
LR	0.78	0.78	0.78	0.78
RF	0.78	0.77	0.77	0.77
MNB	0.80	0.80	0.80	0.80
SVM	0.80	0.79	0.79	0.79
CNN (F)	0.23	0.48	0.31	0.33
CNN (G)	0.24	0.48	0.31	0.34
BiLSTM (F)	0.28	0.51	0.36	0.38
BiLSTM (G)	0.27	0.51	0.35	0.37
CNN + BiLSTM (F)	0.29	0.49	0.36	0.38
CNN + BiLSTM (G)	0.29	0.48	0.36	0.37
Indic-BERT	0.81	0.82	0.81	0.81
m-BERT	0.83	0.81	0.82	0.82
XLM-R	0.86	0.85	0.86	0.85
MuRIL-BERT	**0.88**	**0.87**	**0.87**	**0.87**

Table 4: Performance of employed models on the test set, where F, G, and G-Score represent FastText, GloVe embeddings, and geometric mean score of precision and recall.

Among the ML models, MNB demonstrated an F1-score of 0.80, surpassing both LR (0.78) and SVM (0.79) in overall performance. This outcome indicates that MNB is better suited for this specific task, likely due to its efficiency in handling textual data distributions. Concerning DL models, CNN (G) and CNN (F) achieved F1 scores of 0.31, showing room for improvement in generalization. However, the hybrid CNN+BiLSTM models demonstrated a more robust performance. CNN+BiLSTM (F) achieved an F1 score of 0.36, outperforming both CNN (G) and CNN (F). This improvement highlights the strength of combining CNN's feature extraction capability with BiLSTM's sequential learning ability. However, CNN+BiLSTM (G) yielded a comparable performance with an F1-score of 0.36, slightly underperforming CNN+BiLSTM (F).

Transformer-based models significantly outperformed traditional and deep learning models because they efficiently process contextual information. Indic-BERT and m-BERT achieved F1-scores of 0.81 and 0.82, respectively, demonstrating strong performance for multilingual tasks. XLM-Roberta (XLM-R) further improved with an F1-score of 0.86, showcasing its capability in handling large-scale contextual information across diverse linguistic settings. Finally, MuRIL-BERT outperformed all other models, achieving the highest F1-score of 0.87 and G score of 0.87. The superior performance of MuRIL-BERT can be attributed to its robust contextual understanding, fine-tuned hyperparameters, and optimal training over nine epochs. This analysis highlights the consistent superiority of transformer-based models, particularly MuRIL-BERT, underscoring their ability to generalize well to multilingual and complex datasets. Appendix B presents a detailed error analysis of the proposed model's performance in detecting fake news in Malayalam. MuRIL-BERT performed well in detecting fake news from Malayalam social media texts due to its multilingual pretraining with a strong focus on Indian languages, including Malayalam. Unlike general multilingual models, MuRIL is trained on monolingual and transliterated text, allowing it to capture language-specific patterns common in social media. Fine-tuning domain-specific fake news data further enhanced its contextual understanding, enabling it to differentiate between misinformation cues, sentiment shifts, and propaganda techniques. This combination of pretraining advantages, contextual awareness, and careful optimization contributed to MuRIL-BERT achieving the best results in our experiments.

6 Conclusion

This work addressed the shared task by exploring various ML, DL, and transformer-based baselines for fake news detection in Malayalam. The results demonstrated that transformer-based models significantly outperformed others, with MuRIL-BERT achieving the highest F1-score of 0.87, demonstrating its superior capability to capture contextual information in multilingual datasets. Future work could explore advanced transformer architectures, such as GPT or ELMo, and integrate contextualized embeddings to enhance performance. Additionally, ensemble approaches that combine multiple transformer models or hybrid architectures tailored for fake news detection could offer even better results by leveraging the strengths of diverse models.

Limitations

The current work on fake news detection has several drawbacks, influenced by the following factors:

- Since the dataset is limited, the model's generalization is not guaranteed.
- Despite leveraging transformer-based models for contextual understanding, the system still struggles with detecting nuanced misinformation, such as subtle propaganda, satire, or region-specific deceptive narratives.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Sharal Coelho, Asha Hegde, Kavya G, and Hosahalli Lakshmaiah Shashirekha. 2023. MUCS@DravidianLangTech2023: Malayalam fake news detection using machine learning approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 288–292, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9(61):1871–1874.

Salman Farsi, Asrarul Eusha, Ariful Islam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech EACL2024: Fake news detection in Malayalam language leveraging fine-tuned MuRIL BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 173–179, St. Julian's, Malta. Association for Computational Linguistics.

Sunil Gundapu and Radhika Mamidi. 2021. Transformer based automatic covid-19 fake news detection system. *Preprint*, arXiv:2101.00180.

R L Hariharan, Mahendranath Jinkathoti, P Sai Prasanna Kumar, and M Anand Kumar. 2024. Fake news detection in telugu language using transformers models. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6.

Chaitra K Hiramath and G. C Deshpande. 2019. Fake news detection using deep learning techniques. In *2019 1st International Conference on Advances in Information Technology (ICAIT)*, pages 411–415.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Sheng How Kong, Li Mei Tan, Keng Hoon Gan, and Nur Hana Samsudin. 2020. Fake news detection using deep learning. In *2020 IEEE 10th Symposium on Computer Applications Industrial Electronics (ISCAIE)*, pages 102–107.

Viera Maslej Krešňáková, Martin Sarnovský, and Peter Butka. 2019. Deep learning methods for fake news detection. In *2019 IEEE 19th International Symposium on Computational Intelligence and Informatics and 7th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics (CINTI-MACRo)*, pages 000143–000148.

Sachin Kumar, Rohan Asthana, Shashwat Upadhyay, Nidhi Upreti, and Mohammad Akbar. 2020. Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, 31(2):e3767. E3767 ETT-19-0216.R1.

Sudhanshu Kumar and Thoudam Doren Singh. 2022. Fake news detection on hindi news dataset. *Global Transitions Proceedings*, 3(1):289–297. International Conference on Intelligent Engineering Approach(ICIEA-2022).

Rania M. San Ahmed. 2021. Fake news detection in low-resourced languages "kurdish language" using machine learning algorithms. 12:4219–4225.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Momina Qazi, Muhammad U.S. Khan, and Mazhar Ali. 2020. Detection of fake news using transformer model. In *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–6.

Mina Schütz, Alexander Schindler, Melanie Siegel, and Kawa Nazemi. 2021. Automatic fake news detection with pre-trained transformer models. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 627–641, Cham. Springer International Publishing.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. Combating hostility: Covid-19 fake news and hostile post detection in social media. *Preprint*, arXiv:2101.03291.

Hrithik Majumdar Shibu, Shrestha Datta, Md. Sumon Miah, Nasrullah Sami, Mahruba Sharmin Chowdhury, and Md. Saiful Islam. 2025. From scarcity to capability: Empowering fake news detection in low-resource languages with llms. *Preprint*, arXiv:2501.09604.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and

Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Yingjia Zhao and Xin Tao. 2021. ZYJ123@DravidianLangTech-EACL2021: Offensive language identification based on XLM-RoBERTa with DPCNN. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 216–221, Kyiv. Association for Computational Linguistics.

A Dataset Visualization

Despite leveraging transformer-based models for contextual understanding, the system still struggles with detecting nuanced misinformation, such as subtle propaganda, satire, or region-specific deceptive narratives.

Figure A.1 represents the most common words in fake news in the training set, which could indicate sensational language. In contrast, Figure A.2 shows the prominent words in the original news in the training set, reflecting the typical vocabulary of factual reporting. This analysis generated word clouds to visualize the most frequent words in fake and original news articles. A maximum of 200 words were used for visualization in each word cloud, with word size proportional to frequency.

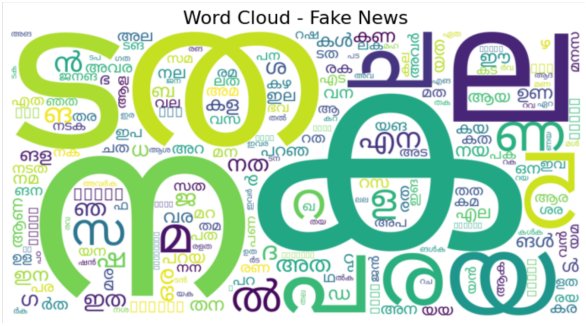


Figure A.1: Word Cloud distribution of *Fake* class.

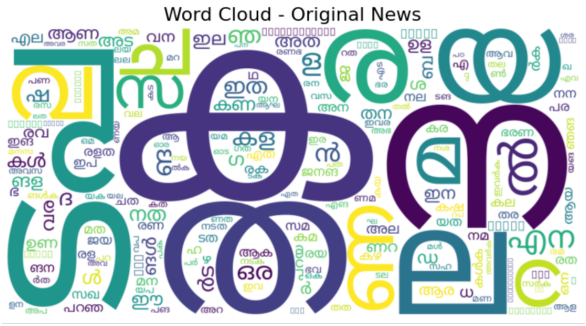


Figure A.2: Word Cloud distribution of *Original* class.

B Error Analysis

We have performed both quantitative and qualitative error analysis to obtain in-depth insights into the performance of the proposed model.

Quantitative Analysis: The MuRIL-BERT was used to conduct a quantitative error analysis, utilizing the confusion matrix shown in Figure B.1. The confusion matrix for the fake news detection

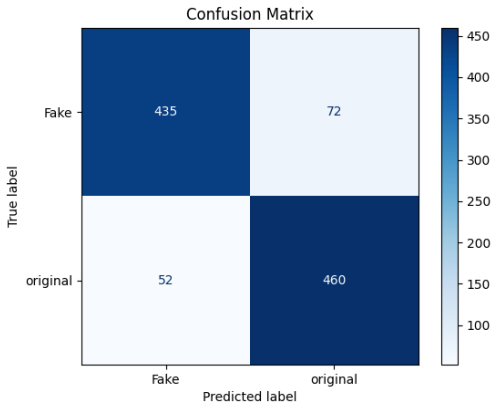


Figure B.1: Confusion matrix of the proposed model (MuRIL-BERT) for fake news detection.

task highlights the classification performance of the proposed MuRIL-BERT model. The model successfully classified most samples, with 435 instances of *Fake* and 460 instances of *Original* being correctly predicted. However, there were misclassifications: 72 *Fake* instances were misclassified as *Original*, while 52 *Original* instances were misclassified as *Fake*. These errors suggest that while the model performs well overall, challenges remain in distinguishing subtle differences between the *Fake* and *Original* categories, likely due to the dataset's overlapping linguistic patterns or contextual ambiguities. Further fine-tuning or incorporating additional contextual cues might improve the model's handling of such edge cases.

Qualitative Analysis: Figure B.2 depicts a qualitative analysis of the predictions made by the proposed MuRIL-BERT model for the fake news detection task. The model successfully classified samples 1 and 5 as *Fake* and samples 3 and 4 as *Original*, which aligns with their respective labels, showcasing its ability to identify a range of text samples correctly. However, the model incorrectly predicted sample 2 as *Original* instead of *Fake*, potentially due to linguistic nuances or overlapping features in the dataset.

Sample Text	Actual Label	Predicted Label
Sample-1: ചേട്ടാ വാർത്ത വയ്ക്കുന്നത് കേരളത്തിലാണ് സംഘി ഭരിക്കുന്ന നോർത്ത് ഇന്ത്യയിലല്ല,ഇവിടെ ആരോഗ്യ മന്ത്രി ഷൈലടീച്ചറാണ്	Fake	Fake
Sample-2: കൊറോണ സിപിഎം നേയും dyfi. യേയും ഭയക്കുന്നു. ബക്കറ്റിൽ പൈസയിടാൻ കാശില്ലാത്തതിനാൽ ഒളിച്ചോടാൻ തയാറെടുക്കുന്നു.	Fake	Original
Sample-3: തിരുവാതിര കളി നടക്കുമ്പോൾ ഗം ഓർത്തു ചിരിച്ചത് ഞാൻ മാത്രമാണോ?? 🤣	Original	Original
Sample-4: മന്ദബുദ്ധികളെ ഭരണഘടന സംരക്ഷിക്കുവാൻ ചുമതലപെടുത്തിയാൽ ഇങ്ങനെയൊക്കെ ഇരിക്കും	Original	Original
Sample-5: അവസരം നൽകൂ, ഏതെങ്കിലും വാദം ഉന്നയിക്കുമ്പോഴേക്കും ജയിലിൽ ഇടാൻ നോക്കുന്നതിനു എന്തിനാണ് , IMO പറയുന്നതിൽ വല്ല കാര്യവുമുണ്ടോ എന്നറിയേണ്ടേ	Fake	Fake

Figure B.2: Some predicted outputs by the proposed method (MuRIL-BERT).

LexiLogic@DravidianLangTech 2025: Detecting Fake News in Malayalam and AI-Generated Product Reviews in Tamil and Malayalam

Souvik Bhattacharyya*, **Pranav Gupta***, **Niranjan Kumar M, Billodal Roy**
Lowe's
Correspondence: {souvik.bhattacharyya, pranav.gupta, niranjan.k.m, billodal.roy}@lowes.com

Abstract

Fake news and hard-to-detect AI-generated content are pressing issues in online media, which are expected to exacerbate due to the recent advances in generative AI. Moreover, tools to keep such content under check are less accurate for languages with less available online data. In this paper, we describe our submissions to two shared tasks at the NAACL Dravidian Language Tech workshop, namely detecting fake news in Malayalam and detecting AI-generated product reviews in Malayalam and Tamil. We obtained test macro F1 scores of 0.29 and 0.82 in the multi-class and binary classification sub-tasks within the Malayalam fake news task, and test macro F1 scores of 0.9 and 0.646 in the task of detecting AI-generated product reviews in Malayalam and Tamil respectively.

1 Introduction

The proliferation of AI-generated content and misleading content such as fake news has resulted in concerns in multiple domains such as e-commerce, news and social media, and other digital domains. Existing tools to detect such content have been restricted to high-resource languages. Moreover, it is challenging for language models to learn complex and rapidly evolving trends and cultural attributes that determine whether something is false or misleading.

With the advent of large language models (LLMs) like GPT (Radford et al., 2018) and Llama (Touvron et al., 2023), generating machine-produced product reviews has become easier than ever. This has huge potential of misleading consumers into purchasing items they might not otherwise choose. As AI tools continue to advance, distinguishing between machine-generated and human-authored text is becoming increasingly challenging. Watermarking LLM output is a novel approach to mitigating the spread of LLM-generated text on the internet, whether in the form of product reviews, fake news, or propaganda on social media. However, due to the lack of consensus among major corporations, ethical concerns, and the availability of open-weight models, watermarking is no longer a consistently viable solution. This underscores the need to explore alternative approaches to address the growing challenge of detecting machine-generated content.

In this paper, we describe our submissions to 2 tasks at the Dravidian Language Tech workshop at NAACL 2025, namely fake news detection in Malayalam and detecting AI-generated product reviews in Tamil and Malayalam. The fake news task had 2 sub-tasks: performing binary classification of textual news as "original" or "fake," and performing multi-class classification of textual news as "half true," "partly false," "mostly false," and "false". On the test dataset, our submissions ranked **6th** out of 16th and **12th** out of 21 submissions in the binary and multi-class classification sub-tasks respectively.

The task on detecting AI-generated product reviews in Tamil and Malayalam consisted of product reviews with binary labels "human" (human generated) and "AI" (AI-generated). Our submissions ranked **4th** among 51 teams and **30th** among 54 teams in Malayalam and Tamil respectively.[1]

2 Related Work

While fake news detection efforts in Malayalam have been limited, several studies have addressed fake news detection in social media platforms using deep learning techniques (Shu et al., 2017; Ghosh and Mitra, 2017; Dhar and Agarwal, 2018; Subramanian et al., 2025). For low-resource languages like Malayalam, approaches such as transfer learning by fine-tuning multilingual BERT based mod-

*These authors contributed equally to this work.

[1]The code for this work is available at https://github.com/prannerta100/naacl2025-dravidianlangtech.

els (Devlin et al., 2019a; Dabre et al., 2022) have shown promise in earlier shared tasks.

As LLMs become more prevalent, research on detecting AI-generated content has grown. GPTZero (Habibzadeh, 2023) is one such tool developed to address concerns about academic plagiarism, using metrics like perplexity and burstiness, though it has been criticized for its false positive rate. Luo et al., 2023 introduced a supervised learning approach for detecting AI-generated reviews by categorizing linguistic features and training classifiers like kNN, AdaBoost, and SVM. Studies by Kirchenbauer et al., 2024 have focused on watermarking LLM outputs by embedding statistical patterns into machine-generated text that can still be detected algorithmically despite alterations like token replacements or paraphrasing. In contrast, DetectGPT, proposed by Mitchell et al., 2023, avoids the need for a separate classifier or explicit watermarking. Instead, it calculates *perturbation discrepancy* using log probabilities from the model of interest and random perturbations applied to the passage, checking if this discrepancy exceeds a predefined threshold. More recent work by Bahad et al., 2024 adopts OpenAI's approach of fine-tuning a RoBERTa-based model (Liu et al., 2019) on a diverse dataset which demonstrated strong performance in identifying the source language model among multiple candidates.

3 Fake News Detection

3.1 Binary Classification

In this task, we were given a dataset (Subramanian et al., 2024, 2023; Devika et al., 2024) of news in Malayalam, with labels "original" and "fake." The dataset consisted of social media posts in pure and code-mixed Malayalam, both in English and Malayalam scripts.

The details of the dataset are given in Table 1

Label	Train set	Dev set	Test set
Original	1658	409	512
Fake	1599	406	507

Table 1: Dataset details for the binary classification subtask

We tried the following 4 models for this subtask:

1. **TFIDF + Logistic Regression:** TFIDF vectorization was a popular method for creating features out of textual data before the advent of foundational neural language models such as BERT. Moreover, logistic regression on top of TFIDF features provides a simple, linear baseline with lesser chances of overfitting. We used the default 'scikit-learn' parameters for training the binary classifier, with a maximum solver iteration parameter of 100.

2. **Fasttext:** Fasttext (Joulin et al., 2016) is a library for efficient learning of word representations and text classification. It uses shallow neural networks for text classification, and includes other in-built optimizations for efficient model training.

3. **GPT-4o:** GPT-4o is an instruction-finetuned large language model by OpenAI, used for a variety of NLP applications. We used GPT-4o with selected training examples and the following prompt: *(system) You are an NLP expert helping classify Malayalam fake news. Before outputting, you will think what the text means within the cultural context of a Malayalam speaker.*
(user) You are a classifier. Use the training data below to classify each text as 'original' or 'fake', output only a json that is a list of records with fields 'text' and 'prediction':

4. **Malayalam BERT**: Malayalam BERT is a monolingual BERT model trained from publicly available monolingual Malayalam datasets (Joshi, 2022). We finetuned this BERT model for the binary classification subtask, given the model's ability to understand Malayalam text. Larger multilingual models are harder to finetune, hence we can expect Malayalam BERT to capture the nuances of Malayalam fake news better. We trained the model for 5 epochs on the train dataset, while using a learning rate of 2×10^{-5}, a weight decay regularization parameter of 0.01, and a per-device batch size of 32.

Table 2 summarizes the performance of our models we tried for this sub-task. We see that Malayalam BERT outperforms the other models. While Fasttext achieves a similar test accuracy as Malayalam BERT, the train-test performance gap is much higher, indicating overfitting to the train set. Such overfitting is not observed in the finetuned Malayalam BERT.

Model	Train F1	Test F1
TFIDF + Log. Reg.	0.928	0.769
FastText	0.9957	0.805
GPT-4o	-	0.782
`malayalam-bert`	0.851	**0.808**

Table 2: Train and Test set Macro F1 scores for the binary classification sub-task

3.2 Multi-class Classification

In this task, we were given a dataset (Subramanian et al., 2024, 2023; Devika et al., 2024) of news in Malayalam, with labels "half true", "partly false", "mostly false", and "false." The dataset consisted of social media posts in pure and code-mixed Malayalam, both in English and Malayalam scripts.

The details of the dataset are given in Table 3.

Label	Train set	Test set
FALSE	1386	100
MOSTLY FALSE	295	56
HALF TRUE	162	37
PARTLY FALSE	57	7

Table 3: Dataset details for the multi-class classification sub-task

We see that the train and test datasets have significant data imbalance, with "partly false" entries being roughly an order of magnitude less frequent. This makes classification more challenging, given the lack of cases that can teach the model a clear distinction between minority and other classes. We tried the same models as the binary classification sub-task with the same hyperparameters. The only exception was our prompt for GPT-4o, which was different from the binary classification sub-task. The GPT-4o is described below:

(system) You are an NLP expert helping classify fake news in Kerala. Before outputting, you will think what the text means within the cultural context of a Malayali. The categories like false, half true, etc. will tell how trustworthy the news text is. For example, 'half true' means the text is half true. Follow reasoning like this:

1. Think about what this sentence means, and put in the larger societal context of Kerala.

2. Revisit the training examples, and check whether your prediction agrees with the kind of labels that the training examples have.

3. Make sure you choose your final answer after carefully weighing the possibilities, for example, is it 'mostly false' or 'false'.

(user) You are a classifier. Use the training data below to classify each text as [FALSE, MOSTLY FALSE, HALF TRUE, PARTLY FALSE], output only a json that is a list of records with fields 'text' and 'prediction':

Table 4 summarizes the performance of our models we tested for this sub-task. We see that GPT-4o performed the best, and TFIDF + Logistic Regression did better than Fasttext and Malayalam BERT, a surprising result. However, the train-test performance gap is higher for logistic regression. This needs further exploration, as a better choice of hyperparameters combined with synthetic data or undersampling the majority classes might help improve the test set macro F1. We experimented with synthetic data generated by GPT-4o, but it did not yield noticeable performance improvements.

Model	Train Macro F1	Test Macro F1
TFIDF + Log. Reg.	0.297	0.203
FastText	0.209	0.167
`malayalam-bert`	0.209	0.167
GPT-4o	-	**0.290**

Table 4: Train and Test set Macro F1 scores for the multi-class classification sub-task

4 Detecting AI-generated product reviews

4.1 Dataset and Task Description

The dataset provided by the organizers for sub-task 5 (Premjith et al., 2025) was divided into separate subsets for Tamil and Malayalam. Each dataset comprised a mix of human-generated and machine-generated product reviews. The objective of this task is to develop and evaluate models capable of accurately distinguishing between AI-generated and human-generated reviews in these languages, effectively addressing a binary classification problem. The distribution of classes is shown below.

Language	Human Gen.	AI Gen.
Malayalam	400	400
Tamil	403	405

Table 5: Class Distribution for Tamil and Malayalam Datasets

4.2 Methods

We adopted a fine-tuning approach using several transformer-based encoder and decoder models. Our base models included the multilingual BERT base model (Devlin et al., 2019b), two monolingual BERT models released by L3Cube (Joshi, 2022), and GPT-2 (Radford et al., 2019). The multilingual BERT model was pre-trained on 102 languages with masked language modeling (MLM) and next sentence prediction (NSP) objectives. The monolingual BERT models were fine-tuned from the existing multilingual model using a monolingual corpus. For GPT-2, we utilized the 124M parameter version model, a transformer-based decoder-only model pre-trained on a large dataset with a causal language modeling (CLM) objective.

The fine-tuning process involves initializing each model with pre-trained weights and adding a classification head on top with 10% dropout. For each subtask, we fine-tune the entire model on the given dataset using stochastic gradient descent with back-propagation. As the subtask is a binary classification problem, we use cross-entropy loss as the objective function. Each dataset is split into 80% for training and 20% for testing, and we utilize the ADAM optimizer (Kingma and Ba, 2017) with an exponential learning rate scheduler. Table 6 shows the training hyperparameters we used.

Parameter	**Value**
Learning rate	5e-5
Learning rate decay	0.9
Batch size	32
Training epochs	10

Table 6: Training hyperparameters

4.3 Results

We fine-tuned all three models on the training set for 10 epochs. During training, we tracked accuracy, precision, F1 scores, and training loss. In our experiment `bert-base-multilingual`, `tamil-bert`, and `malayalam-bert` outperformed GPT-2 in all their respective tasks.

Tables 7 and 8 show the observed results for the Tamil and Malayalam language task respectively.

Tables 9 shows the performance of our submitted models and overall ranks on the held-out test set. The difference in macro F1 scores between our test set and the held-out test set suggests that the final test set included more complex and subtle texts, which posed greater challenges for our fine-tuned model to identify effectively.

Metric	bert-base-multi	tamil-bert	gpt2
Accuracy	0.9938	0.9877	0.9568
Precision	0.9885	0.9773	0.9438
Recall	1.0000	1.0000	0.9767
F1-Score	**0.9942**	0.9885	0.9600

Table 7: Evaluation of fine-tuned models on Tamil test set

Metric	bert-base-multi	malayalam-bert	gpt2
Accuracy	0.9688	0.9688	0.9062
Precision	0.9870	0.9630	0.9012
Recall	0.9500	0.9750	0.9125
F1-Score	0.9682	**0.9689**	0.9068

Table 8: Evaluation of fine-tuned models on Malayalam test set

Language	Macro F1	Rank
Malayalam	0.9	4/51
Tamil	0.646	30/54

Table 9: Evaluation of fine-tuned models on held-out test set

5 Conclusion

In this paper we explore modeling approaches for detecting fake news in Malayalam, a low-resource Dravidian language, and tackle the challenge of identifying AI-generated product reviews in Malayalam and Tamil.

For detecting fake news in Malayalam, we explored several approaches within the binary and multi-class classification sub-tasks. We discovered that while simpler approaches like Fasttext yield good performance on the test set, the overfitting is much higher than fine-tuned transformer models such as `malayalam-bert`. In multi-class classification, we were unable to achieve significant macro F1s and saw that GPT-4o did the best, indicating the need for further exploration and error analysis.

For detecting AI-generated product reviews we tested several transformer-based models, including multilingual and monolingual BERT models and GPT-2, fine-tuning them on provided datasets. Our results showed that BERT-based models outperformed GPT-2 in most cases.

6 Limitations

While our paper compares various models and discusses their ability in detecting fake news and AI-generated product reviews in Dravidian languages, there are certain limitations as well. Fake news on social media and AI-generated content in e-commerce are rapidly evolving issues, which are more difficult to detect in lower-resource languages and for communities with complex cultural nuances and rapidly evolving social landscapes. These challenges are further compounded by the continuous changes in language use and the emergence of new forms of disinformation, making it essential for models to be adaptive to shifting patterns. Approaches such as active and continual learning and other qualitative feedback mechanisms are necessary to combat such issues, as they enable the system to update its knowledge base and improve its accuracy over time. Furthermore, real-world challenges such as domain adaptation present additional difficulties; models trained on one domain may not perform optimally when transferred to another, due to differences in context, vocabulary, and cultural references. Moreover, such NLP-based systems might be biased towards or against certain views, thus unintentionally suppressing the opinions of well-meaning individuals. These biases can emerge due to the data used to train models, which may not be representative of diverse viewpoints or communities. The lack of diverse training data can inadvertently lead to the marginalization of certain demographic groups. Larger unsupervised and supervised datasets are necessary to capture such nuances, in order to avoid socioeconomic biases in online platforms using models described in this paper. Additionally, the presence of these biases can affect the real-world applicability of these models, as they may produce skewed results when deployed in different contexts or for different populations. Furthermore, given the black-box nature of the models in our paper, we also need to focus on investigating their interpretability and explainability. Understanding how these models arrive at their decisions is crucial for addressing potential biases and improving their fairness, as well as for fostering trust among users and stakeholders.

References

Sankalp Bahad, Yash Bhaskar, and Parameswari Krishnamurthy. 2024. Fine-tuning language models for AI vs human generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 918–921, Mexico City, Mexico. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Arpita Dhar and Anjali Agarwal. 2018. Fake news detection using deep learning models. In *Proceedings of the 2018 ICML*, volume 97, pages 1008–1018. PMLR.

Arpita Ghosh and Pabitra Mitra. 2017. Detecting fake news in social media: A data mining perspective. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 743–752. ACM.

Farrokh Habibzadeh. 2023. GPTZero performance in identifying artificial intelligence-generated medical texts: A preliminary study. *J. Korean Med. Sci.*, 38(38):e319.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2024. A watermark for large language models. *Preprint*, arXiv:2301.10226.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Jiwei Luo, Guofang Nan, Dahui Li, and Yong Tan. 2023. Ai-generated review detection. *Available at SSRN 4610727*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *Preprint*, arXiv:2301.11305.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, Kumaresan Thavareesan, Sajeetha, and Prasanna Kumar. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kai Shu, Anna Sliva, Siyi Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 1–10. SIAM.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

SSNTrio @ DravidianLangTech 2025: Hybrid Approach for Hate Speech Detection in Dravidian Languages with Text and Audio Modalities

Bhuvana J
Sri Sivasubramaniya Nadar College of Engineering
bhuvanaj@ssn.edu.in

Mirnalinee T T
Sri Sivasubramaniya Nadar College of Engineering
MirnalineeTT@ssn.edu.in

Rohan R
Sri Sivasubramaniya Nadar College of Engineering
rohan2210124@ssn.edu.in

Diya Seshan
Sri Sivasubramaniya Nadar College of Engineering
diya2210208@ssn.edu.in

Avaneesh Koushik
Sri Sivasubramaniya Nadar College of Engineering
avaneesh2210179@ssn.edu.in

Abstract

This paper presents the approach and findings from the Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) shared task at DravidianLangTech@NAACL 2025. The task focuses on detecting multimodal hate speech in Tamil, Malayalam, and Telugu, requiring models to analyze both text and speech components from social media content. The proposed methodology uses language-specific BERT models for the provided text transcripts, followed by multimodal feature extraction techniques, and classification using a Random Forest classifier to enhance performance across the three languages. The models achieved a macro-F1 score of 0.7332 (Rank 1) in Tamil, 0.7511 (Rank 1) in Malayalam, and 0.3758 (Rank 2) in Telugu, demonstrating the effectiveness of the approach in multilingual settings. The models performed well despite the challenges posed by limited resources, highlighting the potential of language-specific BERT models and multimodal techniques in hate speech detection for Dravidian languages.

1 Introduction

The rapid growth of social media has led to an increase in online hate speech, making automated detection a crucial task for maintaining safe digital spaces. Hate speech refers to content that promotes hate, discrimination, or offensive remarks, while non-hate speech encompasses neutral or non-offensive content. The presence of multimodal content- combining text, speech, and other media- adds complexity to this problem, particularly in underrepresented languages. The Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) shared task at DravidianLangTech@NAACL 2025 focuses on multimodal hate speech detection in Tamil, Malayalam, and Telugu, presenting unique challenges due to linguistic diversity and resource limitations.

This shared task aims to develop robust models that can analyze textual as well as speech components of social media content and classify them accordingly (Premjith et al., 2024a). The task evaluates models based on their Macro Average F1 score, a common metric used in NLP to measure the performance of classification models. The datasets for Tamil, Malayalam, and Telugu pose distinct challenges, such as variations in script, phonetics, and contextual interpretation of hate speech (Sreelakshmi et al., 2024).

This paper details the methodology used to address these challenges, incorporating language-specific BERT models, followed by multimodal feature extraction, and classification using a Random Forest classifier. The results highlight the effectiveness of the approach in handling multimodal hate speech detection while also emphasizing the challenges in lower-resource languages like Telugu (Premjith et al., 2024b). The findings provide insights into improving multimodal learning for Dravidian languages and contribute to the broader field of hate speech detection in multilingual social media contexts.

2 Related Works

The detection of hate speech in social media has garnered significant attention in recent years due to the growing concerns surrounding online harassment and toxicity. Early studies primarily focused on text-based hate speech detection, employing traditional machine learning techniques such as support vector machines (SVM), random forests, and naive bayes (El-Sayed et al., 2023).

However, with the rise of deep learning, researchers began to explore neural network-based approaches for automatic feature extraction and classification. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) showed promise for handling text classification tasks, especially sentiment and hate speech analysis (Kumar, 2022). More recently, transformer-

based models like BERT have further advanced the field, enhancing text classification performance through contextualized word embeddings (Saleh et al., 2021).

In the context of multimodal hate speech detection, the inclusion of audio, visual, and text modalities has been explored to improve classification accuracy. These approaches are especially effective in identifying subtle forms of hate speech, where non-verbal cues and tone of speech are crucial. For instance, work on hate speech detection in video content has incorporated both speech recognition and computer vision techniques to capture the audio and visual aspects of hate speech (Das et al., 2023).

Despite advancements in multimodal hate speech detection, challenges persist, especially in low-resource languages like Tamil, Malayalam, and Telugu. While some progress has been made through language-specific models and dataset augmentation (Azam et al., 2022), hate speech detection in these languages is still under-researched. The presented work uses language-specific BERT models and multimodal feature extraction to improve performance for these languages.

3 Dataset Description

The dataset provided for this task consists of text and speech components sourced from social media platforms (Lal G et al., 2025). The dataset has been curated to reflect real-world social media discourse, ensuring a diverse representation of linguistic patterns, phonetic variations, and hate speech expressions in Tamil, Malayalam, and Telugu.

The Tamil, Malayalam, and Telugu train datasets consist of 514, 883, and 556 rows respectively. The test datasets of the 3 languages consist of 50 rows each.

Each data sample comprises:

- **Text Modality**: Transcribed text extracted from social media posts, incorporating code-mixed language, informal expressions, and slang commonly used in online communication.
- **Speech Modality**: Audio samples corresponding to spoken content, covering diverse accents, intonations, and pronunciations specific to each Dravidian language.

The dataset is annotated for hate speech classification, with labels indicating the presence or absence of hate speech. The labeling schema categorizes hate speech based on its type and severity across three languages: Malayalam, Tamil, and Telugu. Each language dataset includes two main classes - Hate and Non-Hate (N). The Hate class is further divided into four subclasses: Gender (G), Political (P), Religious (R), and Personal Defamation (C). Detailed dataset statistics and description are provided in Table 1.

Labels	Tamil	Malayalam	Telugu
G	68	82	106
R	61	91	72
P	33	118	58
C	65	186	122
N	287	406	198

Table 1: Distribution of Hate Speech Labels

4 Methodology

4.1 Dataset Preprocessing

The first step in the methodology involves preprocessing both the text and audio data. For the text data, tokenization is applied, stop words are removed, and the text is cleaned by eliminating unnecessary characters, such as punctuation and special symbols. The audio data undergoes noise reduction using spectral subtraction to improve quality and ensure consistency. In spectral subtraction, the noise spectrum is estimated from silent or low-energy portions of the audio signal, and this is subtracted from the speech signal. This results in faster training and improved accuracy, leading to more reliable predictions in tasks such as hate-speech detection.

4.2 Data Upsampling

To address the class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to upsample the minority class. SMOTE generates synthetic samples by interpolating between existing minority class instances and their nearest neighbors. This process balances the dataset, ensuring that the classifier is not biased towards the majority class. By incorporating these synthetic samples, the model's ability to effectively learn from both classes is enhanced, improving overall classification performance. All minority classes were upsampled to an equivalent count to match that of the majority class, ensuring a more balanced representation across all classes.

4.3 Feature Extraction

In the feature extraction phase, the preprocessed text is passed through a language specific BERT model such as Tamil BERT, Malayalam BERT or Telugu BERT, depending on the sub task, which generates embeddings that provide contextualized word representations. These embeddings are then used as the feature set for the classifier. Subsequently, features are extracted from the audio using techniques including Mel-Frequency Cepstral Coefficients (MFCC) and spectral representations, which capture key auditory information. These multimodal features, representing both textual and auditory aspects of the data, are utilized as additional input features for the classification task.

4.4 Model Building

After feature extraction, the text and audio features were combined and tested with various classification models, including Random Forest, Support Vector Machine (SVM), and other suitable algorithms, to predict the target labels based on the multimodal features.

Among these classifiers, Random Forest yielded the best results due to its ability to handle high-dimensional feature spaces by combining multiple decision trees, which reduces overfitting and improves generalization. It naturally performs feature selection, focusing on the most relevant data, further enhancing its performance.

Additionally, Random Forest is more robust compared to SVM, which requires careful parameter tuning for varying feature scales. Its ability to manage feature importance and adapt to diverse data made it the most effective model for this task.

5 Result Analysis

The performance of the model was evaluated using standard metrics, including accuracy, precision, recall, and F1-score. The results for each metric are shown in Table 2, which outlines the model's performance across different evaluation criteria.

The proposed model demonstrated strong performance across multiple low-resource languages, achieving macro-F1 scores of 0.7332 (Rank 1) in Tamil, 0.7511 (Rank 1) in Malayalam, and 0.3758 (Rank 2) in Telugu. Figures 1-3 illustrate the confusion matrix for all the three models. From these matrices, it is evident that the confusion matrices for Tamil and Malayalam are quite similar.

The model's strong macro-F1 scores for both

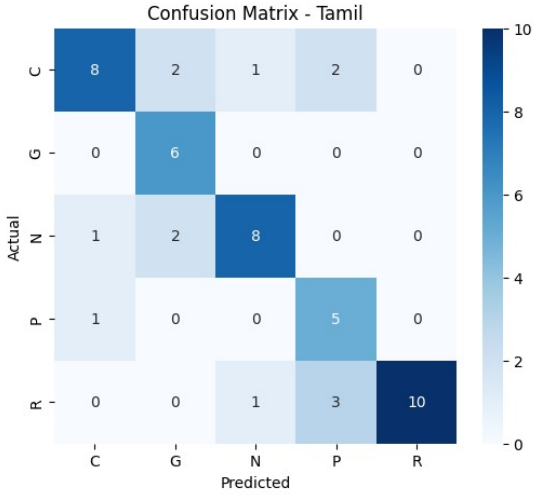


Figure 1: Confusion Matrix for Tamil Dataset

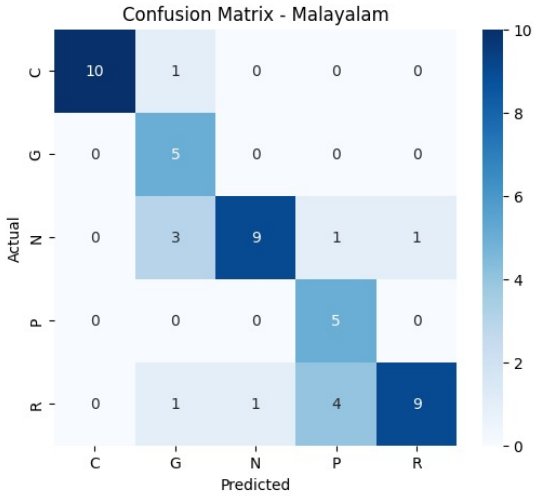


Figure 2: Confusion Matrix for Malayalam Dataset

Tamil and Malayalam suggest that the combination of BERT for text processing, MFCC-based audio features, and Random Forest for feature fusion was highly effective in capturing the linguistic and acoustic nuances of these languages. The high macro-F1 scores indicate that the model was able to balance precision and recall effectively in the presence of class imbalances.

However, the macro-F1 score for Telugu was notably lower compared to Tamil and Malayalam, which could be attributed to differences in pronunciation, phonetics, and linguistic patterns. The model can further be enhanced using advanced techniques and better fine-tuning in order to improve its performance for Telugu.

Overall, these findings demonstrate the efficacy of a multimodal approach for detecting hate speech in low-resource languages, showing that it can sur-

Language	Precision	Recall	F1 Score	Accuracy
Tamil	0.78	0.74	0.73	0.74
Malayalam	0.83	0.76	0.75	0.76
Telugu	0.43	0.36	0.38	0.36

Table 2: Performance Metrics

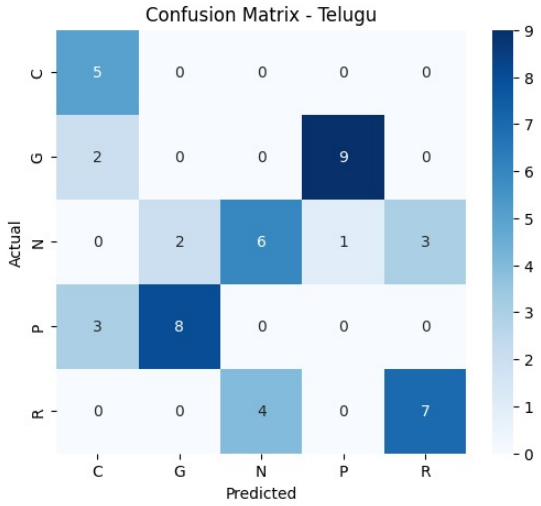


Figure 3: Confusion Matrix for Telugu Dataset

pass conventional text-only techniques by utilizing complimentary data from both modalities.

6 Conclusion

In conclusion, the proposed multimodal model for hate speech detection, which integrates both audio and text inputs, offers a promising approach for improving performance in low-resource languages. By leveraging the unique features of both modalities, this approach improves the model's ability to effectively identify hate speech, especially in underrepresented languages. The proposed method uses language-specific BERT models for text and traditional audio extraction techniques like Mel-Frequency Cepstral Coefficients (MFCC), allowing for a more robust detection process.

The results highlight a significant improvement in performance when combining audio and text, demonstrating the potential of multimodal approaches in detecting hate speech within resource-constrained environments. Overall, our approach presents a reliable and efficient solution for detecting hate speech in low-resource languages, paving the way for future advancements in the field of multimodal hate speech detection.

7 Future Enhancements

For future enhancements, one promising direction is the fine-tuning of models with domain-specific data. Although the current models have been trained on general datasets, focusing on more specific domains or social media platforms could provide more context-aware models for detecting hate speech, especially in niche topics.

Multi-modal fusion techniques also offer an exciting avenue for further enhancement. Standard techniques are already used to combine text and audio features, but exploring more sophisticated fusion strategies such as early, late, or hybrid fusion could lead to better integration of these modalities. This could improve the model's ability to effectively capture the interaction between text and speech, particularly in situations where one modality (e.g., audio) complements the other (e.g., text).

8 Limitations

While the proposed multimodal model demonstrates improved performance in hate speech detection for low-resource languages, several limitations must be considered. First, the effectiveness of the model is highly dependent on the availability and quality of labeled datasets for both text and audio modalities. Many low-resource languages lack sufficiently large and diverse datasets, which can hinder the model's generalizability.

Additionally, the reliance on language-specific BERT models may introduce biases if the pretraining data does not adequately represent different dialects, variations, or informal speech patterns. In the audio modality, challenges such as background noise, variations in pronunciation, and differences in recording quality can affect feature extraction techniques like Mel-Frequency Cepstral Coefficients (MFCC), potentially leading to misclassifications.

Moreover, multimodal models require higher computational resources compared to unimodal approaches, making real-time deployment and scalability in resource-constrained environments challenging. The fusion of audio and text features also

introduces complexities in feature alignment, particularly when dealing with asynchronous or incomplete data inputs.

Furthermore, interpretability remains a concern, as transformer-based models and deep learning approaches often function as black-box systems, making it difficult to provide clear explanations for classification decisions. Addressing these limitations through improved data augmentation, noise-robust feature extraction, and explainability techniques will be crucial for enhancing the effectiveness and practicality of multimodal hate speech detection models.

References

Ubaid Azam, Hammad Rizwan, and Asim Karim. 2022. Exploring data augmentation strategies for hate speech detection in Roman Urdu. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4523–4531, Marseille, France. European Language Resources Association.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. *Preprint*, arXiv:2305.03915.

Tharwat El-Sayed, Abdallah Mustafa, Ayman El-Sayed, and Mohamed Elrashidy. 2023. Hate speech detection by classic machine learning. In *2023 3rd International Conference on Electronic Engineering (ICEEM)*, pages 1–4.

Anuj Kumar. 2022. A study: Hate speech and offensive language detection in textual data by using rnn, cnn, lstm and bert model. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1–6.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2021. Detection of hate speech using bert and hate speech word embedding with deep model. *Preprint*, arXiv:2111.01515.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

Fired_from_NLP@DravidianLangTech 2025: A Multimodal Approach for Detecting Misogynistic Content in Tamil and Malayalam Memes

Md. Sajid Alam Chowdhury, Mostak Mahmud Chowdhury, Anik Mahmud Shanto, Jidan Al Abrar, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1904{064, 055, 049, 080}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

In the context of online platforms, identifying misogynistic content in memes is crucial for maintaining a safe and respectful environment. While most research has focused on high-resource languages, there is limited work on languages like Tamil and Malayalam. To address this gap, we have participated in the Misogyny Meme Detection task organized by DravidianLangTech@NAACL 2025, utilizing the provided dataset named MDMD (Misogyny Detection Meme Dataset), which consists of Tamil and Malayalam memes. In this paper, we have proposed a multimodal approach combining visual and textual features to detect misogynistic content. Through a comparative analysis of different model configurations, combining various deep learning-based CNN architectures and transformer-based models, we have developed fine-tuned multimodal models that effectively identify misogynistic memes in Tamil and Malayalam. We have achieved an F1 score of 0.678 for Tamil memes and 0.803 for Malayalam memes.

1 Introduction

The rapid proliferation of social media has enabled the widespread sharing of memes, which are often used to express humor, ideas, or opinions. However, this medium is also increasingly being misused to propagate harmful ideologies, including misogyny. Therefore, detecting misogynistic content in memes has become essential for mitigating hate speech and ensuring online safety.

Many works have been done on harmful meme detection (Sharma et al., 2022), (Lin et al., 2024), (Gu et al., 2024), (Pramanick et al., 2021), but only a limited number of studies have specifically focused on misogyny meme detection (Srivastava, 2022), (Fersini et al., 2019), (Habash et al., 2022). Most of the existing research in misogyny detection has concentrated on high-resource languages like English, Hindi, and Arabic (Singh et al., 2024), (Srivastava, 2022), (Mulki and Ghanem, 2021), (Mahdaouy et al., 2022), leveraging large-scale datasets and advanced techniques and models. However, research in low-resource languages such as Tamil and Malayalam has been scarce (Rajalakshmi et al., 2023), (Ghanghor et al., 2021), (Chakravarthi et al., 2024), leaving a significant gap in addressing this issue in multilingual and diverse online communities.

Figure 1: Example of a misogynistic and a non-misogynistic meme in Tamil

To address this gap, the task of Misogyny Meme Detection was introduced as part of DravidianLangTech@NAACL 2025. For this task, the organizers have provided a dataset named MDMD (Misogyny Detection Meme Dataset) for memes in the Tamil and the Malayalam languages (Ponnusamy et al., 2024), consisting of both misogynistic and non-misogynistic memes. The details of this shared task and its findings have been thoroughly presented in the overview paper (Chakravarthi et al., 2025).

Our objective has been to develop a multimodal model that effectively detects misogynistic memes in Tamil and Malayalam by combining both visual and textual elements. To achieve this, we have employed various CNN-based architectures and transformer-based models and conducted a comparative analysis of different multimodal model configurations.

Our main contributions are as follows:

- We have developed fine-tuned multimodal models that can effectively detect misogynistic memes in Tamil and Malayalam.
- We have conducted a comparative analysis of various model configurations, combining different transformer-based models with CNN backbones.

The implementation details are available in this GitHub repository[1].

2 Related Work

Recent research has focused on multimodal approaches for detecting harmful content in memes, particularly misogynistic and offensive memes.

Several studies have proposed frameworks that fuse both text and image features to improve detection accuracy. For instance, the MISTRA framework has been developed by utilizing variational autoencoders for dimensionality reduction of image features and combining them with text embeddings to detect misogynous memes (Jindal et al., 2024). In (Pramanick et al., 2021), the authors have introduced MOMENTA, a multimodal deep neural network that analyzes both global and local perspectives within memes to detect harmful content. Additionally, a large-scale Hindi-English code-mixed dataset has been introduced in (Singh et al., 2024), focusing on misogynous meme detection using multimodal fusion methods.

The authors in (Gu et al., 2024) have proposed the SCARE framework, which addresses multimodal alignment by maximizing the mutual information between image and text features while enhancing intra-modal representation learning. Furthermore, in (Habash et al., 2022), an ensemble of models has been utilized by combining multiple multimodal deep learning models for detecting misogynous content.

In the field of multilingual meme detection, the DravidianLangTech-2022 shared task in (Das et al., 2022) has explored meme detection in Tamil, showing that fusing text-based and image-based models improves performance for troll meme classification. The authors in (Ghanghor et al., 2021) have focused on offensive language identification and troll meme classification in multiple Dravidian languages. Moreover, in (Chakravarthi et al., 2024), an overview of the first shared task on 'Multitask Meme Classification - Unraveling Misogynistic and Troll Memes in Online Memes' has been presented, focusing on Tamil and Malayalam memes.

[1] https://github.com/Sajid064/Misogyny-Meme-Detection

3 Dataset

The Misogyny Meme Detection task of DravidianLangTech@NAACL 2025 consisted of two subtasks: one for the Tamil language and the other for the Malayalam language. We were provided with the MDMD dataset, which contains memes and text transcriptions for each language, annotated as either misogynistic or non-misogynistic (Ponnusamy et al., 2024).

Tamil Dataset			
Category	**Train**	**Dev**	**Test**
Non-Misogyny	851	210	267
Misogyny	285	74	89
Total	1136	284	356

Table 1: Dataset distribution for Tamil memes.

Malayalam Dataset			
Category	**Train**	**Dev**	**Test**
Non-Misogyny	381	97	122
Misogyny	259	63	78
Total	640	160	200

Table 2: Dataset distribution for Malayalam memes.

Tables 1 and 2 present the dataset distribution for the Tamil and Malayalam languages, respectively. The Tamil language consisted of 1,336 training samples, 284 validation samples, and 356 test samples, while the Malayalam consisted of 640 training samples, 160 validation samples, and 200 test samples.

We can see that the dataset is highly imbalanced, with non-misogynistic memes significantly outnumbering misogynistic ones. Additionally, many text transcriptions have contained code-mixed text, combining English with Tamil or Malayalam. The images have also included redundant elements such as social media logos, profile names, and icons.

4 Methodology

This section presents our approach for misogyny meme detection in Tamil and Malayalam. The methodology consists of four main components: input modalities, preprocessing, feature extraction, and cross-modal attention and fusion. Figure 2 summarizes the model architecture.

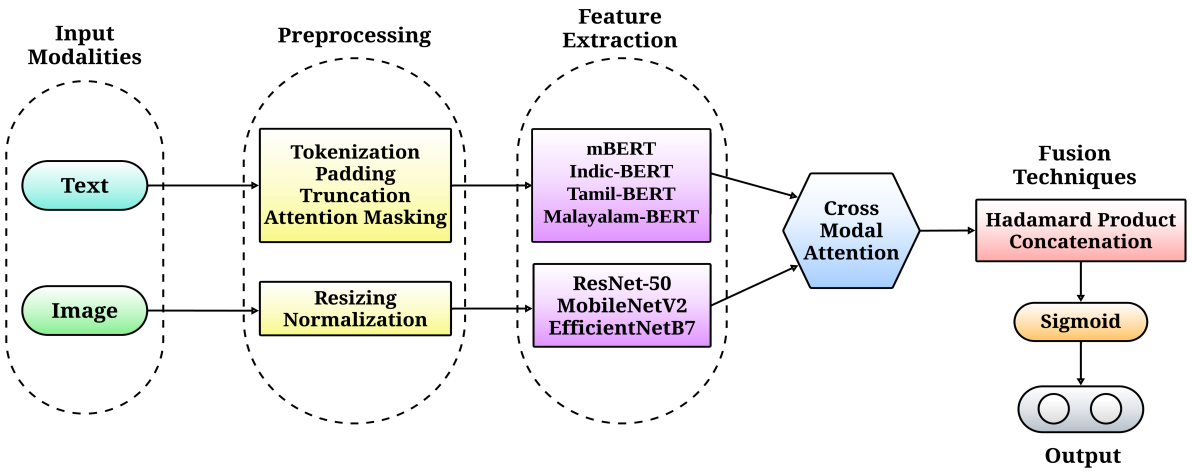


Figure 2: Model architecture of our proposed multimodal approach for misogynistic meme detection

4.1 Input Modalities

We have utilized two primary input modalities:

- **Text Modality:** Textual data has been obtained from transcriptions and processed using a transformer-based language model.
- **Image Modality:** Images corresponding to the textual descriptions have been processed using a deep convolutional neural network (CNN) backbone.

4.2 Preprocessing

4.2.1 Text Processing

The text data have been pre-processed using the BERT tokenizer to convert raw text into input token sequences. We have used padding to ensure that all sequences are of uniform length. We have truncated the sequences if they exceed the maximum length. Then, we applied attention masking to distinguish real tokens from padding tokens. The processed tokens, including the attention mask, have then been fed into a pre-trained Tamil BERT model.

4.2.2 Image Processing

The images have been resized to 224×224 pixels for consistency and to match the input requirements and normalized for faster convergence.

4.3 Feature Extraction

4.3.1 Text Feature Extraction

We have utilized multiple transformer-based models for text feature extraction. For both Tamil and Malayalam datasets, we have employed two general-purpose multilingual models (mBERT and Indic-BERT) to ensure robust multilingual representations. Additionally, we have used two language-specific BERT models: Tamil-BERT for Tamil texts and Malayalam-BERT for Malayalam texts. Each input text has been tokenized and passed through the transformer-based models, and we have extracted the `pooler_output` representation from the final transformer layer.

4.3.2 Image Feature Extraction

For image-based feature extraction, we have experimented with multiple deep CNN architectures, including ResNet50, MobileNetV2, and EfficientNetB7. These models have been initialized with ImageNet pre-trained weights, and their fully connected layers have been removed to obtain meaningful feature representations.

4.4 Cross-Modal Attention and Fusion

To effectively combine textual and visual information, a cross-modal attention mechanism has been applied so that the model can focus on the most relevant aspects of both modalities by computing attention scores between text and image features. For fusion, we have employed both the Hadamard product and concatenation techniques. The Hadamard product has been used for element-wise interaction between the attended image and text features, and the concatenated representation has been used to preserve distinct modality-specific characteristics. Finally, the fused features have then been passed through dense layers for final classification.

5 Experimental Setup

The parameter setups for our multimodal model are displayed in Table 3.

Parameter	Value
Optimizer	Adam
Loss Function	Binary Crossentropy
Learning Rate	$1e^{-4}$
Learning Rate Scheduler	Factor: 0.5
	Patience: 3
	Min lr: $1e^{-7}$
Early Stopping	Patience: 10
Batch Size	8
Epochs	100

Table 3: Training Parameter Settings

6 Experimental Findings

In this section, we have provided the experimental results of our proposed model. Table 4 shows a comparative analysis of different model configurations, combining various BERT variants with CNN backbones by evaluating the Micro-F1 score on the test samples of the Tamil (TAM) and Malayalam (MAL) datasets.

BERT Variants	CNN Backbone	F1 Score	
		TAM	MAL
mBERT	ResNet50	0.621	0.710
	MobileNetV2	0.596	0.681
	EfficientNetB7	0.644	0.742
Indic-BERT	ResNet50	0.573	0.717
	MobileNetV2	0.566	0.694
	EfficientNetB7	0.598	0.728
Tamil-BERT	ResNet50	0.647	-
	MobileNetV2	0.658	-
	EfficientNetB7	**0.678**	-
Malayalam -BERT	ResNet50	-	0.794
	MobileNetV2	-	0.773
	EfficientNetB7	-	**0.803**

Table 4: Performance comparison of different models using various BERT variants and CNN backbones

Among the general-purpose multilingual models, we have observed that the mBERT model consistently outperforms the Indic-BERT model across all the CNN backbones used (ResNet50, MobileNetV2, and EfficientNetB7). The combination of mBERT and EfficientNetB7 has achieved the highest F1 score of 0.644 for Tamil memes and 0.742 for Malayalam memes. In contrast, Indic-BERT with EfficientNetB7 has obtained a lower F1 score of 0.598 for Tamil and 0.728 for Malayalam.

For language-specific models, Tamil-BERT paired with EfficientNetB7 has demonstrated superior performance for Tamil memes by achieving the highest F1 score of 0.678 and surpassing all multilingual models. Similarly, Malayalam-BERT with EfficientNetB7 has achieved the highest F1 score of 0.803 for Malayalam memes by outperforming other configurations. These results indicate that while multilingual models like mBERT have performed well, language-specific models fine-tuned on their respective languages have yielded better results.

7 Error Analysis

From Table 1 and 2, we have observed that the distribution of Tamil memes is highly imbalanced, with a significantly larger number of non-misogynistic samples compared to misogynistic samples. This has led to lower F1 scores as our model has struggled with the minority class. In contrast, the class distribution of Malayalam memes is slightly more balanced, leading to comparatively improved performance. Additionally, the overall dataset size is quite limited, which has restricted the model's ability to generalize effectively. Another major challenge has been the presence of code-mixed text, where many transcriptions have been in English-written Tamil/Malayalam or a combination of English and Tamil/Malayalam words. This has made it harder for BERT models to extract proper features. Furthermore, a significant number of images in the dataset contained redundant elements such as social media icons, profile names, and profile photos, which have introduced noise into the learning process. These distractions have also contributed to some misclassifications.

8 Conclusion

In this paper, we have developed fine-tuned multimodal models for the detection of misogynistic memes in Tamil and Malayalam. Through a comparative analysis of various model configurations, combining transformer-based models with CNN backbones, we have found that language-specific BERT models combined with powerful CNN architectures, such as EfficientNet, achieved the highest results for both languages. In the future, we plan to experiment with more advanced models, such as Vision Transformers (ViT), and explore techniques to mitigate the dataset imbalance issue for enhanced performance.

9 Limitations

While our proposed approach has shown promising results, certain limitations have remained. The availability of labeled data has been limited, which has impacted the ability of our model to generalize effectively to unseen instances. Additionally, the approach has not explicitly accounted for the cultural and linguistic subtleties of the Tamil and Malayalam languages, which may have influenced classification accuracy. Moreover, pre-trained models have inherited biases from their training data, and the multimodal fusion process has faced challenges in capturing implicit or sarcastic expressions of misogyny.

References

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. hate-alert@ dravidianlangtech-acl2022: Ensembling multi-modalities for tamil trollmeme classification. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 51–57.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

Tianlong Gu, Mingfeng Feng, Xuan Feng, and Xuemin Wang. 2024. Scare: A novel framework to enhance chinese harmful memes detection. *IEEE Transactions on Affective Computing*, pages 1–14.

Mohammad Habash, Yahya Daqour, Malak Abdullah, and Mahmoud Al-Ayyoub. 2022. YMAI at SemEval-2022 task 5: Detecting misogyny in memes using VisualBERT and MMBT MultiModal pre-trained models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 780–784, Seattle, United States. Association for Computational Linguistics.

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 2359–2370.

Abdelkader El Mahdaouy, Abdellah El Mekki, Ahmed Oumar, Hajar Mousannif, and Ismail Berrada. 2022. Deep multi-task models for misogyny identification and categorization on arabic social media. *Preprint*, arXiv:2206.08407.

Hala Mulki and Bilal Ghanem. 2021. Working notes of the workshop arabic misogyny identification (armi-2021). *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Faerie Mattins R., Pavitra Vasudevan, and Anand Kumar M.

2023. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech Language*, 78:101464.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar I. Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Y. Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *International Joint Conference on Artificial Intelligence*.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Harshvardhan Srivastava. 2022. Misogynistic meme detection using early fusion model with graph network. *ArXiv*, abs/2203.16781.

One_by_zero@DravidianLangTech 2025: Fake News Detection in Malayalam Language Leveraging Transformer-based Approach

Dola Chakraborty*, **Shamima Afroz***
Jawad Hossain and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904012, u1904106, u1704039}@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

The rapid spread of misinformation in the digital era presents critical challenges for fake news detection, especially in low-resource languages like Malayalam, which lack the extensive datasets and pre-trained models available for widely spoken languages. This gap in resources makes it harder to build robust systems for combating misinformation, despite the significant societal and political consequences it can have. To address these challenges, we propose a transformer-based approach for Task 1 of the Fake News Detection in Dravidian Languages (DravidianLangTech@NAACL 2025), which focuses on classifying Malayalam social media texts as either *original* or *fake*. Our experiments involved a range of machine learning techniques, including Logistic Regression (LR), Support Vector Machines (SVM), and Decision Trees (DT), as well as deep learning architectures such as BiLSTM, BiLSTM-LSTM, and BiLSTM-CNN. Additionally, we explored transformer-based models, including IndicBERT, MuRIL, XLM-RoBERTa, and Malayalam BERT. Among these, Malayalam BERT achieved the best performance, with a macro F1-score of 0.892, securing us a rank of 3rd in the competition.

1 Introduction

Over the past several years, the proliferation of online social media has significantly transformed how individuals communicate, exchange information, and keep up with current affairs (Olan et al., 2024). Platforms such as Twitter, Facebook, and YouTube have enabled users to exchange information at an unprecedented scale (Sharif et al., 2021a). However, this convenience comes with a notable downside: a substantial portion of the information emerging on these platforms is false and, in many cases, intentionally designed to mislead users. Such content, commonly referred to as "fake news," encompasses any false or misleading information presented as original news (Melchior and Oliveira, 2024). The instantaneous reach of social media enables misinformation to spread rapidly, influencing public opinion and causing significant societal, organizational, and political repercussions. While considerable research has been conducted on fake news detection in high resource languages like Spanish, English (Sharma and Singh, 2024; Martínez-Gallego et al., 2021; Hu et al., 2024), low-resource Dravidian languages like Malayalam remain relatively underexplored. Malayalam, spoken mainly in the southern Indian state of Kerala (Thara and Poornachandran, 2022), presents unique linguistic challenges, such as dialect variations, limited annotated datasets and the complex morphology of the language. These aspects make fake news detection in Malayalam a particularly daunting task.

Existing attempts to address fake news in low-resource languages like Malayalam are often constrained by limited datasets, noisy code-mixed data, and the focus of state-of-the-art techniques on high-resource languages. To overcome these limitations, the shared task Fake News Detection in Dravidian Languages(Subramanian et al., 2024), organized by DravidianLangTech@NAACL 2025[1], introduces Task 1 which focuses on classifying social media texts and YouTube comments in Dravidian languages (Malayalam), as either *fake* or *original*.

As participants in this shared task, our work makes the following notable contributions:

- Proposed a transformer-based model specifically designed to classify Malayalam news content as "fake" and "original". This approach harnesses the capabilities of pre-trained language models to effectively tackle the challenges associated with processing

*Authors contributed equally to this work.

[1] https://codalab.lisn.upsaclay.fr/competitions/20698

Malayalam, a low-resource language, in diverse domains such as social media posts and YouTube comments.

- Investigated a range of machine learning, deep learning, and fine-tuned transformer-based architectures to evaluate their effectiveness in detecting fake news and to analyze errors for deeper insights into the detection process.

The implementation of our proposed approach has been made publicly available, and the source code can be accessed on GitHub[2].

2 Related Work

Numerous studies have been conducted on fake news detection, primarily focusing on high-resource languages like English, while paying less attention to low-resource languages like Malayalam. Sharma and Singh (2024) pAhuja and Kumar (2023) proposed Mul-FaD, an attention-based model for fake news detection tested on English, German, and French news articles. The dataset, comprising 43,488 articles, was created by combining English datasets and translating parts into French and German. Mul-FaD achieved the best performance with 93.73% accuracy and an F1 score of 92.9, outperforming baseline models for multilingual fake news detection. Othman et al. (2024) explored Arabic fake news detection using hybrid models combining Arabic pre-trained BERT models (AraBERT, GigaBERT, MARBERT) with CNNs, with AraBERT-2D-CNN achieving the best F1-scores on Arabic datasets ANS (0.6188), Ara-News (0.7837), and Covid19Fakes (0.8009). Rahman et al. (2022) used the BFNC dataset for fake news detection, with XLM-R achieving the best performance, attaining an F1-score of 98% on the test data. Osama et al. (2024) performed the DravidianLangTech@EACL2024 shared task, tackling fake news detection in Malayalam using machine learning, deep learning, and transformer models. Their best model, m-BERT, achieved a macro F1-score of 0.85, ranking 4th and demonstrating its effectiveness in combating misinformation. Rahman et al. (2024) presented the shared task "Fake News Detection in Dravidian Languages - DravidianLangTech@EACL 2024," focusing on identifying fake and original news in Malayalam social media. Their teams employed diverse strategies, from machine learning to transformer models. Malayalam-BERT achieved the best performance with a macro F1-score of 0.88, securing 1st place. Farsi et al. (2024) conducted the DravidianLangTech@EACL2024 shared task focused on detecting fake news in Malayalam. Task 1 involved binary classification (fake or not), and Task 2 was multi-classclassification (five levels). Using machine learning, deep learning, and transformer models, they fine-tuned MuRIL, achieving F1-scores of 0.86 (Task 1) and 0.5191 (Task 2), securing 3rd place in Task 1 and 1st place in Task 2. Bala and Krishnamurthy (2023) implemented the MuRIL base variant model and achieved a notable F1-score of 87% for Malayalam code-mixed text. Balaji et al. (2023) proposed transformer models such as M-BERT, ALBERT, BERT, and XLNET. M-BERT outperformed competitors with a robust F1-score of 0.74, surpassing XLNET and ALBERT, which achieved accuracy scores of 0.71 and 0.66, respectively. Sharif et al. (2021b) presented a detailed description of a system developed for detecting COVID-19 fake news in English (Task-A) and hostile post detection in Hindi (Task-B) using SVM, CNN, BiLSTM, and CNN+BiLSTM with TF-IDF and Word2Vec embeddings. Their system achieved notable results, with the highest weighted F1 score of 94.39% in Task-A and 86.03% coarse-grained and 50.98% fine-grained F1 scores in Task-B. Shyam and Poornachandran (2021) investigated a dataset of Malayalam-English code-mixed text from YouTube comments, evaluating models like Camem-BERT, Distil-BERT, ELECTRA, and XLM-R, with ELECTRA achieving an outstanding F1-score of 99.33%.

3 Task and Dataset Description:

This shared task (Subramanian et al., 2025) focuses on detecting fake news in the Dravidian language Malayalam. Task 1 requires classifying social media texts as either original or fake. The dataset (Devika et al., 2024; Subramanian et al., 2025, 2024, 2023), provided by the organizers, was curated from various social media platforms, including Twitter and Facebook. Table 1 illustrates the distribution of the dataset, which is fairly balanced. The training dataset consists of 1,658 original and 1,599 fake samples. Similarly, the validation dataset contains 409 original and 406 fake samples, while the test dataset includes 512 original and 507 fake sam-

[2] https://github.com/DolaChakraborty12/Fake_News_Detection_Dravidian_Language

ples.

Classes	Train	Valid	Test	T_W
Original	1658	409	512	21626
Fake	1599	406	507	35629
Total	3257	815	1019	57255

Table 1: Dataset Statistics for Train, Validation, and Test Sets. (T_W denotes total words)

4 Methodology

We have implemented various ML, DL, and transformer-based approaches with hyperparameters fine-tuned to find out the best model for this task. Figure 1 depicts a schematic process in detecting fake news, illustrating each major phase.

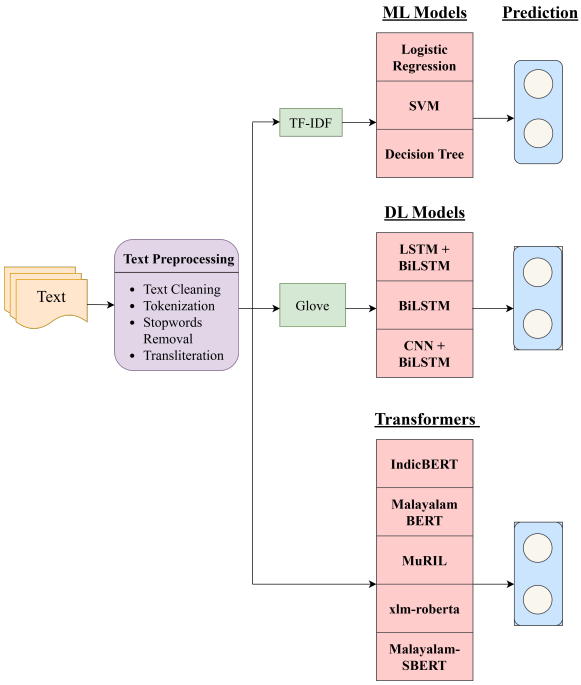


Figure 1: Schematic process of fake news detection.

4.1 Data Preprocessing

To ensure effective training and evaluation of our models, we have implemented an extensive data preprocessing pipeline. The pipeline ensured consistency and clarity in the datasets. The process involved removing punctuation, emojis, special characters, numerical text, and URLs to clean the data. HTML tags were eliminated to retain plain text, and frequent stopwords in both Malayalam and English were filtered out using the Malaya NLP library. Additionally, code-mixed text was standardized by transliterating it into Malayalam using AI4Bharat's transliteration engine. This streamlined preprocessing resulted in a refined dataset suitable for effective model training and evaluation.

4.2 Feature Extraction

Before passing text as input to machine learning, deep learning, and transformer-based models, it was first converted into a numerical format. For feature extraction, we used TF-IDF, GloVe, and Keras. TF-IDF assigns weights based on word frequency, with a vocabulary size of 10,000. GloVe provides pre-trained embeddings with an embedding matrix shape of (10,000, 100). The Keras embedding layer has an input dimension of 10,000 and an output dimension of 128, converting tokenized text into numerical sequences. This approach combines both statistical and semantic representations of text.

4.3 Machine Learning Models

For this task, we explored three machine learning models: Logistic Regression, Decision Trees, and Support Vector Machines (SVM). TF-IDF was employed for feature extraction. The Logistic Regression model utilized a regularization value of 0.01 to mitigate overfitting. The Decision Tree model was configured with a maximum depth of 10, while the SVM model employed a linear kernel for linear classification. Table 2 presents the hyperparameter of the machine learning based models.

Hyperparameter	Value
Max Depth (Decision Tree)	10
Regularization (Logistic Regression)	0.01
random state	42
Maximum Iterations	1000
class weight	balanced
Kernel (SVM)	linear

Table 2: Hyperparameters for machine learning models

4.4 Deep Learning Models

We utilized three deep learning models: BiLSTM, LSTM + BiLSTM, and CNN + BiLSTM, each starting with an embedding layer to process the input text.

- **BiLSTM Model:** This model employed a Bidirectional LSTM layer with 64 units to capture contextual patterns in both forward and backward directions. Dropout layers with rates of 0.8 and 0.5 were added to minimize overfitting.

- **LSTM + BiLSTM Model:** An additional LSTM layer with 64 units (returning sequences) was introduced before the BiLSTM layer. Dropout layers were included alongside an L2-regularized dense layer to enhance generalization. Dropout layer rates is 0.8 and 0.5.

- **CNN + BiLSTM Model:** This model combined a Conv1D layer with 128 filters and a kernel size of 5, followed by a MaxPooling1D layer for feature extraction, before passing the output to a BiLSTM layer with 64 units. To address overfitting, higher dropout rates of 0.9 and 0.8 were applied.

Table 3 presents the hyperparameters of deep learning based models.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	2e-5
Epoch	10
Loss Function	Binary Cross-Entropy
Dropout (BiLSTM, LSTM+BiLSTM)	0.8, 0.5
Dropout (CNN + BiLSTM)	0.9, 0.8
Batch Size (BiLSTM, LSTM + BiLSTM)	16
Batch Size (CNN + BiLSTM)	32

Table 3: Hyperparameters for deep learning-based models

4.5 Transformer Models

Transformers have garnered significant attention in recent years due to their exceptional performance across various NLP tasks. For this task, we explored five pre-trained transformer-based models and fine-tuned them on our dataset to evaluate their effectiveness in this domain:

- MURIL: A multilingual model pre-trained on 17 Indian languages and English. It was fine-tuned with a batch size of 16, a learning rate of 1e-5, a sequence length of 60, and trained for 12 epochs. MURIL has proven to be highly efficient in multilingual tasks.

- IndicBERT: Pre-trained on 12 Indic languages along with English. It was fine-tuned with a batch size of 16, a learning rate of 2e-5, a sequence length of 60, and trained for 10 epochs.

- XLM-RoBERTa: Fine-tuned using the same hyperparameters as IndicBERT, but trained for 15 epochs.

- Malayalam-BERT: Specifically pre-trained on Malayalam text, tailored for tasks in the Malayalam language.

- Malayalam Sentence-BERT: Fine-tuned for sentence-pair tasks, optimized for Malayalam sentence-level tasks.

Table 4 presents the hyperparameters of the transformer models which are optimized through extensive experimentation.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	2e-5
Batch Size	16
Max Length	128
Epochs	15

Table 4: Hyperparameter setup for transformer-based models

5 Result Analysis

Table 5 illustrates the performance of the various ML, DL, and transformer-based models explored on the test dataset. The model's performance was evaluated using the macro F1-score. Transformer-based models, particularly Malayalam BERT, outperformed both ML and DL models, achieving the highest macro F1 score of 0.892. Among DL models, BiLSTM had the best score of 0.782, while the LR model led the ML models with a score of 0.5267. Overall, DL models performed better than ML models, but transformer-based models delivered superior performance overall.

5.1 Error Analysis

The performance of the best performed model is further investigated for in depth understanding of its behaviors using quantitative and qualitative error analysis.

5.1.1 Quantitative Analysis

Figure 2 presents the confusion matrix of the best-performing model, Malayalam BERT. A detailed quantitative error analysis of the fine-tuned Malayalam BERT model is performed based on the confusion matrix. It is evident from the confusion matrix that, out of 1,019 samples, 889 are correctly predicted. The model misclassifies 93 original samples as fake and 37 fake samples as original.

Model	P	R	F1
LR	0.75	0.75	0.75
SVM	0.76	0.76	0.76
DT	0.72	0.63	0.59
LSTM + BiLSTM(K)	0.80	0.80	0.80
BiLSTM(K)	0.80	0.80	0.80
CNN + BiLSTM(K)	0.81	0.81	0.81
LSTM + BiLSTM(G)	0.70	0.65	0.63
BiLSTM(G)	0.71	0.65	0.63
Malayalam BERT	**0.88**	**0.88**	**0.89**
IndicBERT	0.85	0.85	0.85
MuRIL	0.85	0.84	0.84
XLM-R	0.85	0.84	0.84
Malayalam S-BERT	0.86	0.86	0.86

Table 5: Performance of various ML, DL, Transformer-based models on the test set. P (Precision), R (Recall), F1 (macro F1-score)

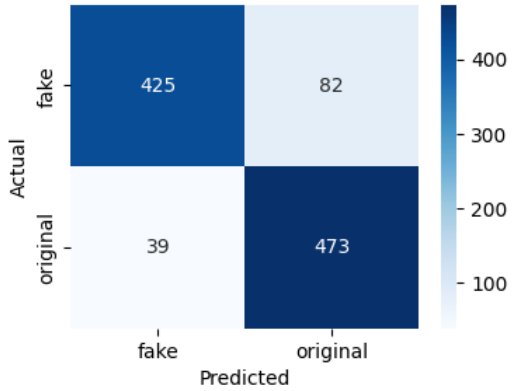


Figure 2: Confusion matrix of Malayalam BERT

5.1.2 Qualitative Analysis

A comparison of actual labels and predicted labels for a particular text is illustrated in Figure 3. The first two samples are incorrectly predicted as original, even though they are fake. However, the next three samples are predicted correctly as their actual classes.The misclassifications likely occur due to the linguistic complexity of Malayalam, including its rich morphology and syntactic structures, which pose challenges for the model in capturing subtle semantic differences. Although Malayalam BERT performs exceptionally well for the Malayalam language, fake news often imitates the style and tone of genuine news, making it challenging for the model to distinguish between the two.

Text	Predicted Label	True Label
5000 ഉള്ള പോൾ ലോഗ്ഡ്വൻ ഇപ്പോള് 250000 എന്താ കാരണം	Original	Fake
ഓഷോ രജനീഷ് പറഞ്ഞപോലെ എനിക്കപ്പോൾ തോന്നിയത് അങ്ങനെയാണ് ..ഇപ്പോൾ തോന്നുന്നത് ഇങ്ങനെയാണ് ...എന്തൊക്കെയോ ആവോ	Original	Fake
ചേട്ടാ വാർത്ത വയ്ക്കുന്നത് കേരളത്തിലാണ് സംഘി ഭരിക്കുന്ന നോർത്ത് ഇന്ത്യയിലല്ല.ഇവിടെ ആരോഗ്യ മന്ത്രി ഷൈലടീച്ചറാണ്	Fake	Fake
Shame for entire Woman's of Kerala	Original	Original
135 code janaghal andhu wide business cheythalum vijayikum in India	Fake	Fake

Figure 3: A few examples of predicted outputs by the Malayalam BERT

6 Conclusion

This paper evaluates the performance of various machine learning, deep learning, and transformer-based models for detecting fake news in Malayalam. While deep learning techniques such as LSTM + BiLSTM, BiLSTM, and CNN + BiLSTM demonstrated strong results, traditional machine learning methods struggled to effectively capture the intricate semantic relationships inherent in the Malayalam language. Among all approaches, Malayalam BERT achieved the best performance, with an F1-score of 0.892, by effectively capturing the language's unique nuances. Future research could focus on enhancing this work by utilizing larger datasets, leveraging ensemble transformer models, and exploring other advanced large language models.

Limitations

Our current work posses some limitations. Some limitations of our work are: i) Malayalam BERT has token limitation, causing truncation of long news articles and potential loss of crucial information. ii) Our task only analyzes text, making it ineffective against misinformation spread via images, memes, or misleading visuals. iii) Due to the small dataset size, the model may struggle to generalize well to diverse fake news patterns.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Nishtha Ahuja and Shailender Kumar. 2023. Mul-fad: attention based detection of multilingual fake news. *Journal of Ambient Intelligence and Humanized Computing*, 14(3):2481–2491.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. AbhiPaw@ DravidianLangTech: Fake news detection in Dravidian languages using multilingual BERT. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–238, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Varsha Balaji, Shahul Hameed T, and Bharathi B. 2023. NLP_SSN_CSE@DravidianLangTech: Fake news detection in Dravidian languages using transformer models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–139, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Salman Farsi, Asrarul Eusha, Ariful Islam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech EACL2024: Fake news detection in Malayalam language leveraging fine-tuned MuRIL BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 173–179, St. Julian's, Malta. Association for Computational Linguistics.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

Kevin Martínez-Gallego, Andrés M Álvarez-Ortiz, and Julián D Arias-Londoño. 2021. Fake news detection in spanish using deep learning techniques. *arXiv preprint arXiv:2110.06461*.

Cristiane Melchior and Mírian Oliveira. 2024. A systematic literature review of the motivations to share fake news on social media platforms and how to fight them. *new media & society*, 26(2):1127–1150.

Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. 2024. Fake news on social media: the impact on society. *Information Systems Frontiers*, 26(2):443–458.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_GoodFellows@DravidianLangTech EACL2024: A transformer-based approach for detecting fake news in Dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192, St. Julian's, Malta. Association for Computational Linguistics.

Nermin Abdelhakim Othman, Doaa S Elzanfaly, and Mostafa Mahmoud M Elhawary. 2024. Arabic fake news detection using deep learning. *IEEE Access*.

MD Sijanur Rahman, Omar Sharif, Avishek Das, Sadia Afroze, and Mohammed Moshiul Hoque. 2022. Fand-x: Fake news detection using transformer-based multilingual masked language model. In *2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 153–158. IEEE.

Tanzim Rahman, Abu Raihan, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@DravidianLangTech EACL2024: Fake news classification using Malayalam-BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 223–228, St. Julian's, Malta. Association for Computational Linguistics.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021a. Combating hostility: Covid-19 fake news and hostile post detection in social media. *arXiv preprint arXiv:2101.03291*.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021b. Combating hostility: Covid-19 fake news and hostile post detection in social media. *CoRR*, abs/2101.03291.

Upasna Sharma and Jaswinder Singh. 2024. A comprehensive overview of fake news detection on social networks. *Social Network Analysis and Mining*, 14(1):120.

Thara Shyam and Prabaharan Poornachandran. 2021. Transformer based language identification for malayalam-english code-mixed text. *IEEE Access*, 9:118837 – 118850.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task

on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

S Thara and Prabaharan Poornachandran. 2022. Social media text analytics of malayalam–english code-mixed using deep learning. *Journal of big Data*, 9(1):45.

CUET_Novice@DravidianLangTech 2025: A Multimodal Transformer-Based Approach for Detecting Misogynistic Memes in Malayalam Language

Khadiza Sultana Sayma, Farjana Alam Tofa, Md Osama and **Ashim Dey**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904013, u1904008, u1804039}@student.cuet.ac.bd, ashim@cuet.ac.bd

Abstract

Memes, combining images and text, are a popular social media medium that can spread humor or harmful content, including misogyny—hatred or discrimination against women. Detecting misogynistic memes in Malayalam is challenging due to their multimodal nature. A Shared Task on Misogyny Meme Detection, organized as part of DravidianLangTech@NAACL 2025, aimed to address this issue by promoting the advancement of multimodal machine learning models for classifying Malayalam memes as misogynistic or non-misogynistic. In this work, we explored visual, textual, and multimodal approaches for meme classification. CNN, ResNet50, Vision Transformer (ViT), and Swin Transformer were used for visual feature extraction, while mBERT, IndicBERT, and MalayalamBERT were employed for textual analysis. Additionally, we experimented with multimodal fusion models, including IndicBERT+ViT, MalayalamBERT+ViT, and MalayalamBERT+Swin. Among these, our MalayalamBERT+Swin Transformer model performed best, achieving the highest weighted F1-score of 0.87631, securing **1st** place in the competition. Our results highlight the effectiveness of multimodal learning in detecting misogynistic Malayalam memes and the need for robust AI models in low-resource languages.

1 Introduction

Misogynistic content fosters hostility and discrimination, particularly targeting specific genders, and poses a significant challenge to creating safe and inclusive online environments. The rise of social media has accelerated the spread of such content, often as text-visual memes. Detecting misogyny in multimodal formats is challenging, as intent depends on text-image interplay. In Malayalam, linguistic complexity adds to the difficulty, requiring advanced tokenization and semantic analysis for effective detection. Addressing these linguistic intricacies is crucial for building robust misogyny detection models. The Misogynistic Meme Detection Shared Task, conducted as part of DravidianLangTech@NAACL 2025 (Chakravarthi et al., 2025), aimed to tackle these challenges by identifying misogynistic content in Tamil and Malayalam memes.

Our participation focused specifically on Malayalam memes. Through this work, we aimed to address the unique challenges posed by the multimodal nature of memes and the intricacies of Malayalam text. Our key contributions include:

- Utilized the Swin Transformer for visual feature extraction, leveraging its advanced capabilities for image representation, and Malayalam-BERT for extracting textual features, given its effectiveness in capturing the nuances of the Malayalam language.
- Additionally, we evaluated the performance of models that were trained exclusively on textual or visual data, which helped us to understand the relative contributions of each modality.

This work not only contributes to the broader goal of misogyny detection in underrepresented languages but also emphasizes the importance of multimodal approaches in tackling the nuanced challenges of meme classification. The code is available at https://github.com/Khadiza13/Misogyny-NAACL2025.

2 Related Work

In recent years, there has been a growing focus among NLP researchers on identifying trolling, hostility, offensive language, and abusive content on social media platforms. While early studies primarily focused on analyzing textual information ((Anzovino et al., 2018) (Sadiq et al., 2021)

(Ishmam and Sharmin, 2019)), recent works have explored multimodal approaches that consider both textual and visual features embedded in memes. (Jha et al., 2024) introduced MultiBully-Ex, a dataset for multimodal explanations in code-mixed cyberbullying memes, combining visual and textual data. Similarly, (Hasan et al., 2022) demonstrated that the CNN-Text+VGG16 combination outperformed other multimodal models with an F1-score of 0.49 for meme detection. (Barman and Das, 2023) utilized mBERT for textual features, ViT for visual features, and MFCC for audio features to tackle abusive language detection. (Rahman et al., 2024) introduced a hybrid ConvLSTM+BiLSTM+MNB model, which obtained the highest macro F1-score of 71.43%. (Ahsan et al., 2024) presented MIMOSA, a new multimodal dataset for detecting Bengali aggression, containing 4,848 annotated memes classified into five aggression categories. They proposed the Multimodal Attentive Fusion (MAF) method. Similarly, (Mahesh et al., 2024) focused on identifying misogynistic memes in Tamil and Malayalam. Their models, including mBERT+ResNet-50 and MuRIL+ResNet-50, obtained macro F1-scores of 0.73 and 0.87, respectively. (Osama et al., 2024) highlighted mBERT's strong performance in misinformation detection for low-resource languages such as Malayalam. Additionally, (Rehman et al., 2025) presented a multimodal framework for detecting misogynistic content using attention, graph-based refinement, and lexicon-based features, achieving notable improvements on benchmark datasets.(Gu et al., 2022) explored ensemble models for misogyny classification, like Naive Bayes and gradient boosting.

3 Task and Dataset Description

A misogynistic meme combines visual content and text to demean, stereotype, or offend women, often spreading harmful ideologies on social media (Ponnusamy et al., 2024). The goal of this task is to classify misogynistic memes by leveraging both visual and textual information (**?**). The organizers provided a dataset containing two types of memes (Misogynistic and Non-misogynistic) in the Tamil and Malayalam languages (Chakravarthi et al., 2024). Here, Table 1 provides the distribution of samples across training, development, and test sets. The dataset is presented as an image accompanied by a corresponding caption.

Class	Train	Dev	Test	Total	Words
Misogynistic	259	63	78	400	3735
Non-misogynistic	381	97	122	600	6560
Total	640	160	200	1000	10295

Table 1: Statistical distribution of our dataset.

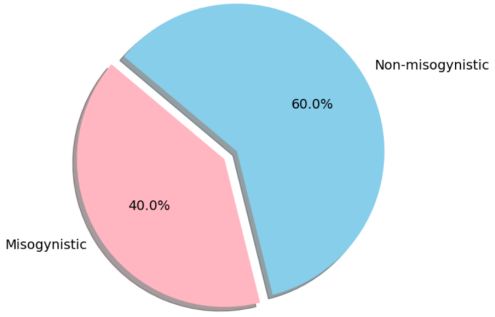


Figure 1: Percentage distribution of two different classes.

Participants can use either the image, the caption, or both to complete the classification task. We employed image, text, and multimodal (image + text) features to tackle the given task.

4 Methodology

The aim of this study is to detect misogynistic content in multimodal Malayalam memes. Initially, we exploit the visual aspects of the memes. Subsequently, the textual information is processed using Malayalam-specific language models, and finally, both modalities are combined through a fusion mechanism to make robust classification decisions. Figure 2 offers a clear visualization of our methodology, highlighting the essential steps in our approach.

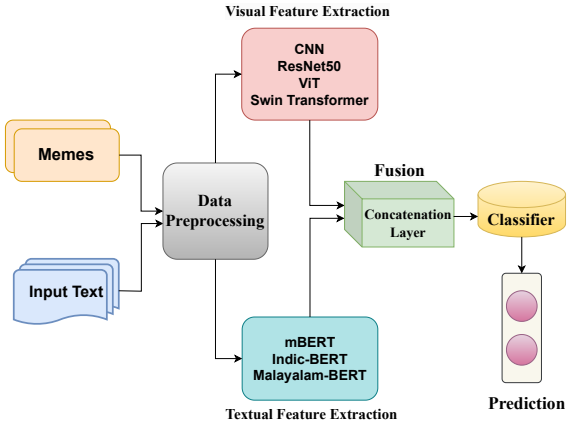


Figure 2: Abstract view of our methodology.

4.1 Data Preprocessing

In this step, the Malayalam text undergoes tokenization using specialized tokenizers from the Malayalam-BERT model. The text is transformed into numerical representations with a maximum sequence length of 128 tokens. Special tokens ([CLS], [SEP]) are added as required by the transformer architecture. For image preprocessing, all memes are transformed into a size of 224×224 pixels, normalized using standard ImageNet statistics (Deng et al., 2009), and converted to RGB format to maintain consistency across the dataset.

4.2 Visual Approach

For visual feature extraction, we first use a CNN with 8 layers, followed by ResNet-50 (He et al., 2016), which is pre-trained on ImageNet. After that, we employ the Vision Transformer (Dosovitskiy et al., 2020) and finally employ the Swin Transformer model (Liu et al., 2021), which utilizes a hierarchical structure with shifted windows for efficient processing of visual information. The model, pre-trained on ImageNet, processes the meme images to generate 1024-dimensional feature vectors. This architecture was chosen for its proven effectiveness in capturing both local and global visual features.

4.3 Textual Approach

The textual component of memes is first processed using BERT-Base Multilingual Cased (Devlin et al., 2018) leveraging pre-trained weights, token resizing, dropout, and a classification layer for meaningful representations. After this, we employ Indic-BERT (Kakwani et al., 2020) and then Malayalam-BERT (Tabassum et al., 2024), a transformer-based model trained for Malayalam, generates 768-dimensional feature vectors, excelling in linguistic nuance and contextual understanding.

4.4 Multimodal Approach

Our multimodal approach combines the visual and textual features through fusion strategy. The visual features from Swin Transformer (1024-dimensional) and textual features from Malayalam-BERT (768-dimensional) are concatenated to form a 1792-dimensional vector. This combined representation is then processed through a two-layer neural network classifier. The first layer reduces the dimensionality to 512, followed by ReLU activation and dropout (0.1) for regularization. The final layer produces binary classification outputs for misogyny detection. The training protocol uses AdamW (learning rate: 5e-5, batch size: 16) for 5 epochs, using binary cross-entropy loss, a learning rate scheduler with warmup steps, and gradient clipping for stability. Table 2 shows the list of tuned hyperparameters used in the experiment.

Hyperparameter	Value
Batch Size	16
Learning Rate	5e-5
Optimizer	AdamW
Epochs	5
Dropout Rate	0.1
Weight Decay	0.01

Table 2: Overview of optimized hyper-parameters.

5 Results & Discussion

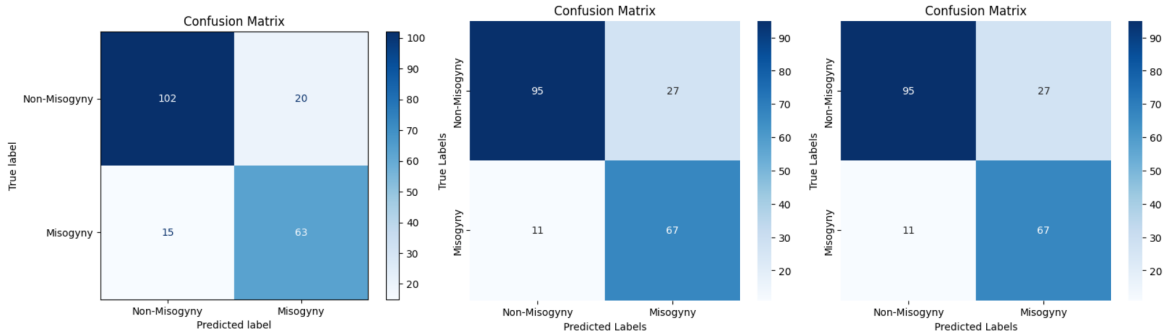
This section presents a comparative performance analysis of various experimental approaches for classifying memes. The effectiveness is primarily assessed based on the weighted f1-score, while precision and recall are also considered in some cases. Table 3 presents a summary of the precision (P), recall (R), and F1 (f1) scores for each model on the test set. The results show that, Swin Transformer and Malayalam-BERT performed best in visual and textual models, respectively, with an F1-score of 0.78. However, the top classification performance was seen in the multimodal models, where combining Malayalam-BERT and Swin Transformer resulted in the highest F1-score of 0.88. These findings highlight the superiority of multimodal models in meme classification by combining text and visual features.

Approach	Classifier	P	R	f1
Visual	CNN	0.62	0.51	0.56
	ResNet50	0.91	0.13	0.22
	ViT	0.87	0.68	0.76
	Swin Transformer	0.76	0.81	0.78
Textual	mBERT	0.69	0.58	0.63
	Indic-BERT	0.62	0.82	0.71
	Malayalam-BERT	0.71	0.86	0.78
Multimodal	Indic-BERT+ViT	0.73	0.83	0.78
	Malayalam-BERT+ViT	0.78	0.81	0.80
	Malayalam-BERT+Swin	**0.87**	0.78	**0.88**

Table 3: Performance of different models on test set.

5.1 Quantitative Discussion

The results underscore the effectiveness of transformer-based architectures in identifying misogynistic content. The confusion matrix in



(a) Visual (b) Texual (c) Multimodal

Figure 3: Confusion matrix of three different approaches.

Figure 3 provides a detailed breakdown of our model's performance. Here, the visual model correctly classifies 102 Non-Misogyny and 63 Misogyny samples but misclassifies 20 instances of Non-Misogyny as Misogyny and 15 Misogyny instances as Non-Misogyny. The textual model improves classification, correctly predicting 95 Non-Misogyny and 67 Misogyny samples, though it misclassifies 27 Non-Misogyny samples. The multimodal model (Malayalam-BERT + Swin Transformer) outperforms both unimodal models, correctly classifying 113 Non-Misogyny and 61 Misogyny instances, with fewer misclassifications (9 false positives and 17 false negatives). These results affirm that multimodal models improve precision and recall in detecting misogynistic memes.

5.2 Qualitative Discussion

Figure 4 presents sample predictions from our best-performing Malayalam-BERT+Swin Transformer model. In the first instance, the model incorrectly classified the sample as non-misogynistic (label 0). This misclassification might have occurred because the text, although seemingly neutral, could have contained subtle contextual cues or sarcasm that the model failed to pick up on. In contrast, the second sample, which was genuinely non-misogynistic (label 0), was misclassified as misogynistic (label 1). This could be attributed to the image associated with the text, which may have included visual elements such as expressions, body language, or symbols that the model interpreted as indicative of misogyny. Furthermore, cultural norms and societal stereotypes in the visual context may have influenced inaccurate predictions. While the model captures linguistic and contextual cues well, it struggles with nuanced cases involving sarcasm, cultural context, or visual ambiguity.

English Translation: To the wife who wakes up in middle of the night putting on makeup Husband: Why are you Wearing makeup at this time??Wife: Thats my phone face lock dude... Now it won't turn on without makeup on

Actual: 1
Predicted: 0

English Translation: WhatsApp status of most people who have seen it for a few days

Actual: 0
Predicted: 1

Figure 4: Examples of some wrongly classified sample of the best model.

6 Conclusion

This work presented the details of the methods and performance analysis of the models for detecting misogynistic memes in Malayalam, exploring visual, textual, and multimodal fusion techniques. The results revealed that the Malayalam BERT+Swin Transformer model got the highest weighted F1-score of 0.88, demonstrating that multimodal fusion significantly enhances model performance. In the future, we plan to explore audio and video modalities, advanced fusion strategies, extend the dataset and ensemble models for better robustness, especially in low-resource languages. Transfer learning, domain-specific knowledge, and addressing social and cultural biases will also enhance the model's adaptability, fairness, and generalization.

Limitations

A primary limitation of this study lies in the reliance on pre-trained models for both visual and textual features, which may not fully capture the nuances of Malayalam-specific cultural context or meme content. While our multimodal approach performs well, the models used are limited by their generalization capabilities when handling domain-specific or low-resource language memes. Additionally, the dataset used for training may not be comprehensive enough to account for all variations in meme content, which could impact the robustness of the model. Furthermore, the impact of cultural norms, humor, and sarcasm—which are often deeply embedded in Malayalam social discourse—has not been explicitly analyzed. Misogynistic content can sometimes be expressed subtly through irony, satire, or culturally specific references, making it difficult for AI models to detect intent accurately. Future work with a larger, more diverse dataset, the exploration of specialized Malayalam language models, and a deeper analysis of sarcasm and cultural factors in error cases could enhance model accuracy and generalizability. While the current dataset was balanced and did not require data augmentation, it would be crucial to incorporate data augmentation techniques when dealing with larger and imbalanced datasets. By generating synthetic examples through text or image transformations, data augmentation could help address class imbalance and improve the model's ability to generalize across different classes. This would ensure better performance, especially in situations where certain classes are underrepresented, ultimately leading to a more robust and reliable model for real-world applications.

References

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M. Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500, St. Julian's, Malta. Association for Computational Linguistics.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.

Shubhankar Barman and Mithun Das. 2023. hatealert@DravidianLangTech: Multimodal abusive language detection and sentiment analysis in Dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 217–224, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Qin Gu, Nino Meisinger, and Anna-Katharina Dick. 2022. Qinian at semeval-2022 task 5: Multi-modal misogyny detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 736–741.

Md Hasan, Nusratul Jannat, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Cuetnlp@ dravidianlangtech-acl2022: Investigating deep learning techniques to detect multimodal troll memes. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 170–176.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 555–560. IEEE.

Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 930–943, St. Julian's, Malta. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Sidharth Mahesh, Sonith D, Gauthamraj Gauthamraj, Kavya G, Asha Hegde, and H Shashirekha. 2024. MUCS@LT-EDI-2024: Exploring joint representation for memes classification. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 282–287, St. Julian's, Malta. Association for Computational Linguistics.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Cuet_nlp_goodfellows@ dravidianlangtech eacl2024: A transformer-based approach for detecting fake news in dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Md. Rahman, Abu Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das, and Mohammed Moshiul Hoque. 2024. Binary_Beasts@DravidianLangTech-EACL 2024: Multimodal abusive language detection in Tamil based on integrated approach of machine learning and deep learning techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 212–217, St. Julian's, Malta. Association for Computational Linguistics.

Mohammad Zia Ur Rehman, Sufyaan Zahoor, Areeb Manzoor, Musharaf Maqbool, and Nagendra Kumar. 2025. A context-aware attention and graph neural network-based multimodal framework for misogyny detection. *Information Processing & Management*, 62(1):103895.

Saima Sadiq, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, Gyu Sang Choi, and Byung-Won On. 2021. Aggression detection through deep neural model on twitter. *Future Generation Computer Systems*, 114:120–129.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_punctuators@ dravidianlangtech-eacl2024: Transformer-based approach for detection and classification of fake news in malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186.

teamiic@DravidianLangTech 2025: Transformer-Based Multimodal Feature Fusion for Misogynistic Meme Detection in Low-Resource Dravidian Language

Harshita Sharma, **Simran**, **Vajratiya Vajrobol**, **Nitisha Aggarwal**,

Institute of Informatics and Communication, University of Delhi, Delhi, India

harshita.sharma@iic.ac.in, simran.2022@iic.ac.in, tiya101@south.du.ac.in, nitisha@south.du.ac.in

Abstract

Misogyny has become a pervasive issue in digital spaces. Misleading gender stereotypes are getting communicated through digital content. This content is majorly displayed as a text-and-image memes. With the growing prevalence of online content, it is essential to develop automated systems capable of detecting such harmful content to ensure safer online environments. This study focuses on the detection of misogynistic memes in two Dravidian languages, Tamil and Malayalam. The proposed model utilizes a pre-trained XLM-RoBERTa (XLM-R) model for text analysis and a Vision Transformer (ViT) for image feature extraction. A custom neural network classifier was trained on integrating the outputs of both modalities to form a unified representation. This model predicts whether the meme represents misogyny or not. This follows an early-fusion strategy since features of both modalities are combined before feeding into the classification model. This approach achieved promising results using a macro F1-score of 0.84066 on the Malayalam test dataset and 0.68830 on the Tamil test dataset. In addition, it is worth noting that this approach secured Rank 7 and 11 in Malayalam and Tamil classification respectively in the shared task of Misogyny Meme Detection (MMD). The findings demonstrate that the multimodal approach significantly enhances the accuracy of detecting misogynistic content compared to text-only or image-only models.

1 Introduction

Misogyny on social media has become an alarming concern in recent years. Unfortunately, digital platforms have become a breeding ground for misogynistic content. This issue is not merely confined to explicit hate speech but extends to more insidious forms like memes that trivialize or normalize sexism. A meme is a cultural idea, joke, trend, or piece of content (often in the form of an image, text, or video) that spreads rapidly through social media. Memes hold the power to shape societal perceptions and reinforce harmful gender stereotypes as these are often humorous, highly engagable, and shareable. Studies have shown that such content exacerbates existing inequalities and fosters a culture of misogyny (Jane, 2017; Banet-Weiser and Miltner, 2015). Addressing this issue is critical for promoting digital civility and safeguarding the rights and dignity of women in online spaces. The detection and mitigation of misogynistic content presents several unique challenges.

Over the years, significant efforts have been made to combat online misogyny. International conventions, policy frameworks, and platform-specific moderation mechanisms are designed. However, the effectiveness of these measures remains limited. Despite advancements in artificial intelligence (AI) and natural language processing (NLP), the detection of misogynistic memes remains an underexplored area. This study focuses on detecting misogynistic memes by addressing their textual and visual complexities. By utilizing a novel dataset provided by the organizers of this shared task this study is focused on extracting robust feature representations using XLM-R (Conneau) and ViT (Dosovitskiy, 2020) embeddings, paired with a carefully designed model architecture to capture nuanced patterns across both classes. The macro-average F1 score ensures balanced evaluation, mitigating class imbalance. The repository is available on GitHub[1]. The goal is to develop models capable of classifying misogynistic content and identifying contextual cues in such memes. Through this research, we seek to contribute to the growing body of knowledge in the field of hate speech detection, advancing the understanding of multimodal misogyny and paving the way for more effective content moderation strategies.

[1]This is link to the code for submission - Sharma (2025)

2 Related Work

The detection of misogynistic content in online platforms has been a growing area of research, particularly in the context of multimodal data such as memes. Early research on misogyny detection and hate speech primarily focused on textual data. (Waseem et al., 2017) highlights NLP-based hate speech detection. With the rise of visual communication, memes have become a prominent medium for spreading hateful content. Zhu et al. (2021) introduced a multimodal approach for detecting visual hate speech in memes, demonstrating that combining textual and image-based features enhances classification performance and Sai et al. (2022) examines different fusion strategies for integrating textual and visual cues showing late fusion approaches. Similarly, Mathew et al. (2021) examined hate speech detection in multilingual settings, highlighting the limitations of unimodal approaches.

Recent research has focused on detecting misogynistic memes in low-resource Dravidian languages, particularly Tamil and Malayalam. Ponnusamy et al. (2024) introduced the MDMD dataset for misogyny detection in Tamil and Malayalam memes, providing a valuable resource for understanding gender bias in these communities. Chakravarthi et al. (2024) organized this shared task, reporting the best macro F1 scores of 0.73 for Tamil and 0.87 for Malayalam. Earlier work by Ghanghor et al. (2021) addressed offensive language identification and troll meme classification in Dravidian languages, achieving weighted F1 scores of 0.75 for Tamil and 0.95 for Malayalam in offensive language detection.

The effectiveness of deep learning in misogyny detection has been well-documented. Johnson and Khoshgoftaar (2019) surveyed deep learning technique in imbalanced class settings. ViTs and CLIP Radford et al. (2021) have shown improvements in multimodal classification performance. Some text-only models (Annamoradnejad and Zoghi, 2024) achieved an F1 score of 0.76, while image-only models (Mathew et al., 2021) reached F1 score of 0.72, demonstrating the limitations of unimodal models. Multimodal models.

Studies such as Sharma et al. (2022) highlight the difficulty of detecting sarcasm and implicit hate speech in multimodal content. Furthermore, Davidson et al. (2017) and Chakravarthi (2022) emphasize the importance of cross-cultural sensitivity in hate speech detection, particularly in multilingual settings. Furthermore, the dataset's inherent bias towards specific cultural and linguistic contexts may limit the model's applicability to global audiences. Singh et al. (2023) employed XLM-RoBERTa to detect hate speech and its targets, demonstrating its effectiveness in multilingual contexts. The Social Media Sexist Content (SMSC) database (Buie and Croft, 2023) and addressing the bias in the dataset (Zhang et al., 2023) provides a valuable benchmark for future research, These studies show progress in multimodal misogyny detection but highlight the need for more research in multilingual, low-resource languages like Tamil and Malayalam.

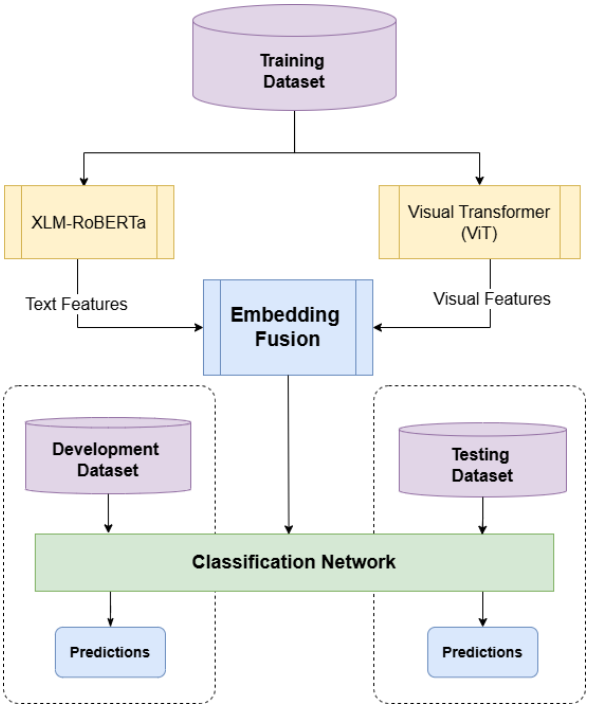


Figure 1: Process Flow of the proposed model for Misogyny Meme Detection.

3 Methodology

The proposed methodology employs a multimodal approach for misogyny detection in Tamil and Malayalam memes using a labeled dataset of images with transcriptions as shown in Figure 1.

3.1 Dataset and data pre-processing

The provided dataset is a unique collection of misogyny memes explicitly focusing on the Tamil and Malayalam languages. Figure 2 represented the transcriptions provided in the dataset (Ponnusamy et al., 2024; Chakravarthi et al., 2024, 2025) include both monolingual and bilingual con-

image_id	labels	transcriptions
1008	0	Sight Adichifying College Staff Expectation Reality nivasmc@fb
1176	1	RUKKu ioh IHL~NW irukkuzingafun irukku PHOTOURID Ungala Leggings Poda venamnu solala Short Tops potutu poda venam nu dha solrom nallavae ila asigama irukum Wear panunga ana Long tops potu wear panunga
76	0	Seven Screen Studio @7screenstudio Considering overflowing passes requests & safety constraints

Figure 2: Attribute of the provided Tamil Training Dataset

image_id	labels	transcriptions
888	0	ഈ ചാടി ഓടി നടക്കണ മനുഷ്യനാണോടാ നിങ്ങളിത്രേം കാലം ബെഞ്ചിലിട്ടത് നീയൊക്കെ പുഴുത്ത് ചാകും വേണ്ടത്. സന്തോഷിച്ചാട്ടെ സന്തോഷിച്ചാട്ടെ.
554	1	മലയാള സിനിമയുടെ ഭാവി വടറാണി ഇവൾ തന്നെ നല്ല കുഴിഞ്ഞ പൊക്കിൾ
556	1	ഒന്ന് പെറ്റത് ആണെങ്കിലും .. മുലയും വയറും ചാടിയതാണെങ്കിലും ചില മലഞ്ചരക്കുകളുടെ ഡിമാൻഡ് ഒരിക്കലും കുറയില്ല

Figure 3: Attribute of the provided Malayalam Training Dataset

tent. Dataset contains code-mixed text that combines Tamil and English. Similarly, Figure 3 combines Malayalam and English languageges. This mix of linguistic elements reflects real-world online discourse and adds complexity to the task.

The dataset was released in three divisions for each language: train, development, and test. Each subset is organized within a folder containing a set of images and an accompanying CSV file with the following structure: 1)image_id: A unique identifier assigned to each image in the folder. 2) labels: Binary annotations indicating whether the image represents misogyny or not. 3)transcriptions: The text transcription is extracted from the corresponding image. In the case of the test dataset, the labels were not provided.

During the initial data analysis, the class distributions in both the Tamil and Malayalam training datasets were examined. The Tamil dataset contains 851 non-misogynistic and 285 misogynistic samples. Malayalam dataset consists of 381 non-misogynistic and 259 misogynistic samples. Instead of applying data balancing techniques such as oversampling or undersampling, the decision was made to retain the original distribution to preserve the natural characteristics of the data. This approach mirrors real-world scenarios, where misogynistic content is generally less prevalent compared to non-misogynistic content (Buie and Croft, 2023). Specifically, the F1-score was used instead of accuracy, as it provides a more reliable measure of performance in cases where class distributions are not perfectly uniform (Johnson and Khoshgoftaar, 2019).

3.2 Preprocessing and Feature Extraction

XLM-R tokenizer was used to tokenised transcriptions of textual data. Each tokenized input was padded and truncated to ensure consistency. The embedding extraction of the text was done by using XLM-R, embeddings were extracted by averaging the last_hidden_state representations of tokens. Raw image files were resized, normalized using ViT. Embeddings were computed using the last_hidden_state of the image model to ensure robust representation of visual features.

3.3 Feature Fusion

The embeddings from text - XLM-R (768 dimensions) and image - ViT (768 dimensions) modalities were concatenated to create a unified representation. Since, here the features from text and images were combined into a single vector and no decision-making occurred at the individual modality level, it can be considered as early-fusion technique. It is a feature-level fusion. This fusion allowed for simultaneous utilization of textual and visual modalities, enhancing the model's ability to classify multimodal inputs effectively.

3.4 Model architecture and training

A custom neural network named MultimodalClassifier, was designed with the following component.
1) Input Layer: Accepts the 1536-dimensional concatenated feature vector.
2) Hidden Layer: A fully connected layer with 512 units and ReLU activation to capture non-linear interactions between the fused features.
3) Output Layer: A fully connected layer that outputs logits corresponding to the number of classes.
The fused embedding was passed through this fully connected neural network classifier. The training process incorporated some model optimizing strategies. CrossEntropyLoss was employed as the loss function. To effectively handle multi-class classification, softmax to the logits was applied. The Adam optimizer (learning rate of 0.001) was used to ensure efficient weight updates. To handle data efficiently, training samples were organized into batches of size 32 using PyTorch's DataLoader. Shuffling was applied to maintain randomness during batch sampling and enhance the model's generalization. The training process was conducted

over ten epochs. For each epoch, a forward pass was executed to compute predictions, followed by a backward pass to calculate gradients by minimizing the loss. The optimizer then updated the model weights. The model's predictions were stored to compute F1-Score along with the loss at the end of each epoch, ensuring a comprehensive evaluation of model performance.

3.5 Evaluation and Metrics

The models went through established evaluation metrics to ensure performance across all classes. In addition to accuracy, the F1-Score (macro) was calculated to provide a holistic view of the model's performance. F1-Score (macro) also gives accurate evaluation for the case of class imbalance. The final predictions of the test dataset in both the languages were evaluated against the macro f1 score metrics.

4 Results and Discussion

The proposed model in this study was evaluated on the provided dataset of misogynistic memes. The results indicate a significant scope of improvement in detecting misogynistic memes when combining both text and image features compared to unimodal approaches, Text-Only (Annamoradnejad and Zoghi, 2024) or Image-Only (Mathew et al., 2021) as it's shown in Sharma et al. (2022) which is also a multimodal approach.

Language	F1-Score	Precision	Recall	Accuracy
Tamil	0.6615	0.6580	0.6659	0.7324
Malayalam	0.7968	0.7964	0.8100	0.8000

Table 1: Performance metrics for Tamil and Malayalam development datasets.

The findings of the shared task indicate satisfactory performance, as reflected in the evaluation metrics for the development dataset. For Tamil and Malayalam, the evaluation metrics including the F1-scores are shown in Table: 1. This research aims to contribute to the expanding field of hate speech detection by enhancing the understanding of multimodal misogyny and informing more effective content moderation strategies. The final results on the test dataset yielded a macro F1-score of 0.84066 for Malayalam language and 0.6883 for Tamil language (see Tables: 3 and 2).

Team	Macro F1-score	Rank
Shraddha	0.70501	10
teamiic	**0.6883**	**11**
InnovationEngineer	0.68782	12

Table 2: F1- score for the Tamil Dataset

Team	Macro F1-score	Rank
CUET-NLP_MP	0.84118	6
teamiic	**0.84066**	**7**
byteSizedLLM	0.83912	8

Table 3: F1- score for the Malayalam Dataset

5 Conclusion

This study demonstrates the effectiveness of a multimodal approach in detecting misogynistic memes, It is highlighting the combining of textual and visual features enhances classification accuracy. By employing XLM-R for text analysis and ViT embeddings for image feature extraction, the model successfully identified nuanced representations of misogynistic content in Tamil and Malayalam memes. The feature fusion strategy significantly contributed to strong classification performance, achieving macro F1-scores of 0.84066 for Malayalam and 0.6883 for Tamil and securing Rank 7 and Rank 11 respectively. These results validate the ability of this approach to tackle the challenges of detecting misogyny in multimodal content. However, the limited size of the labeled data led to the use of pre-trained models like XLM-R and ViT. The reliance of these models on large, diverse datasets for pre-training may hinder their ability to capture highly nuanced, language-specific, or evolving forms of misogyny in the targeted communities. This dependency on pre-trained models may also affect the generalizability of the results to broader online contexts. To overcome these challenges in the future, it may be possible to get a dataset that includes more culturally diverse and real-world samples for enhancing the model's generalizability across different linguistic and societal contexts. As harmful discourse continues to evolve, ongoing advancements in multimodal AI models will play a pivotal role in creating safer, more inclusive online environments.

References

Issa Annamoradnejad and Gohar Zoghi. 2024. Colbert: Using bert sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249:123685.

Sarah Banet-Weiser and Kate M. Miltner. 2015. #Mas-

culinitySoFragile: culture, structure, and networked misogyny. *Feminist Media Studies*, 16(1):171–174.

Hannah Buie and Alyssa Croft. 2023. The social media sexist content (smsc) database: A database of content and comments for research use. *Collabra: Psychology*, 9(1):71341.

Bharathi Raja Chakravarthi. 2022. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

A Conneau. Unsupervised cross-lingual representation learning at scale.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Iiitk@ dravidianlangtech-eacl2021: Offensive language identification and meme classification in tamil, malayalam and kannada. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 222–229.

Emma Jane. 2017. *Misogyny Online: A short (and brutish) history*.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):1–54.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: a benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Siva Sai, Naman Deep Srivastava, and Yashvardhan Sharma. 2022. Explorative application of fusion techniques for multimodal hate speech detection. *SN Computer Science*, 3(2):122.

Dilip Kumar Sharma, Bhuvanesh Singh, Saurabh Agarwal, Hyunsung Kim, and Raj Sharma. 2022. Sarcasm detection over social media platforms using hybrid auto-encoder-based model. *Electronics*, 11(18):2844.

H. Sharma. 2025. teamiic@dravidianlangtech 2025.

Karanpreet Singh, Vajratiya Vajrobol, and Nitisha Aggarwal. 2023. Iic_team@ multimodal hate speech event detection 2023: Detection of hate speech and targets using xlm-roberta-base. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 136–143.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. 2023. Mitigating biases in hate speech detection from a causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6610–6625.

Wanrong Zhu, Xin Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. 2021. Multimodal text style transfer for outdoor vision-and-language navigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1207–1221, Online. Association for Computational Linguistics.

CUET_Novice@DravidianLangTech 2025: Abusive Comment Detection in Malayalam Text Targeting Women on Social Media Using Transformer-Based Models

Farjana Alam Tofa, Khadiza Sultana Sayma, Md Osama and **Ashim Dey**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904008, u1904013, u1804039}@student.cuet.ac.bd, ashim@cuet.ac.bd

Abstract

Social media has become a widely used platform for communication and entertainment, but it has also become a space where abuse and harassment can thrive. Women, in particular, face hateful and abusive comments that reflect gender inequality. This paper discusses our participation in the Abusive Text Targeting Women in Malayalam social media comments for the DravidianLangTech@NAACL 2025 shared task. The task provided a dataset of YouTube comments in Tamil and Malayalam, focusing on sensitive and controversial topics where abusive behavior is prevalent. Our participation focused on the Malayalam dataset, where the goal was to classify comments into these categories accurately. Malayalam-BERT achieved the best performance on the subtask, securing **3rd** place with a macro f1-score of 0.7083, showcasing transformer models' effectiveness for low-resource languages. These results contribute to tackling gender-based abuse and improving online content moderation.

1 Introduction

The rise of social media has changed the way people communicate, share information, and interact with digital content. However, women are frequent targets of abusive comments, including harassment, cyberbullying, and hate speech, which reflect societal biases. Detecting such abuse is crucial for creating safer online spaces. The Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media at DravidianLangTech@NAACL 2025 aims to tackle this challenge. The task focuses on detecting abusive comments targeting women in Tamil and Malayalam, both low-resource languages with challenges like agglutination, rich morphology, and code-mixing. Research on detecting abusive language in low-resource languages, like Tamil and Malayalam has advanced in recent years. The DravidianLangTech shared task (Rajiakodi et al., 2025) introduced benchmark datasets and evaluated transformer-based models for detecting abusive Tamil and Malayalam text targeting women. Their workshop paper (Priyadharshini et al., 2022) presented a dataset for Tamil abusive comment detection. In 2023, another workshop paper (Priyadharshini et al., 2023) introduced datasets for Tamil, Telugu, and code-mixed Tamil-English abusive comment detection. Another paper (Hossain et al., 2022) explored abusive text classification across misogyny, homophobia, and transphobia, addressing dataset imbalances. (Palanikumar et al., 2022) used transliteration-based data augmentation to enhance dataset size and improve model performance in Tamil abusive text detection. Additionally, (M et al., 2023) showed the effectiveness of transformer models for detecting abusive content in multilingual settings. Our participation focused on the Malayalam subtask, where we addressed the complexities of detecting abusive text targeting women. The key contributions of this work are illustrated in the following:

- We explored various ML, DL, and transformer-based models to classify abusive comments in the Malayalam dataset.
- Demonstrated the efficacy of transformer models, including Malayalam-BERT in low-resource languages and advancing the development of content moderation tools.

This work improves abusive language detection for underrepresented languages, fostering safer online platforms. Our code can be accessed at https://github.com/Tofa571/Abusive-Malayalam.

2 Related Work

The detection of abusive language has become a key area of research, especially in low-resource languages. The DravidianLangTech shared task (B et al., 2024) focused on detecting abuse targeting women, where multilingual models outper-

formed language-specific ones. Machine learning approaches such as SVM and SGD (Sivanaiah et al., 2023) addressed Tamil-English code-mixed abuse. It highlighted the significance of addressing class imbalance through undersampling techniques. Another study by (Prithila et al., 2023) introduced a dataset specifically for detecting derogatory comments against women. They emphasized the significance of multilingual datasets and fine-tuning transformer models for improved accuracy. Abusive language detection on social media is challenging due to informal language and limited annotated data in low-resource languages. A co-training framework (Tuarob et al., 2023) utilizes both content and contextual features to improve accuracy, especially for Indic languages. (Zia Ur Rehman et al., 2023) proposed a cross-lingual transformer-based model for Indic languages that incorporates user history and post affinity and shows strong results for low-resource languages like Malayalam. (Sharma et al., 2024) used a CNN-BiLSTM ensemble for gendered abuse detection in Hindi, Tamil, and Indian English but focused on a narrow set of deep learning models, which may limit the ability to handle linguistic nuances in under-resourced settings. Another study (Vetagiri et al., 2024) detects gendered abuse in Hindi, Tamil, and Indian English using a combination of CNN and BiLSTM networks, effectively handling noisy text and code-switching. A dual attention mechanism improved abusive language detection by capturing both internal and contextual relationships, outperforming traditional attention models (Jarquın-Vasquez et al., 2024). The paper (Alharthi et al., 2023) found that online abuse is primarily identity-driven (97%) rather than behavior-driven (3%) and that popular users are more likely to be targeted. In a recent study, (Tofa et al., 2025) evaluated machine learning and transformer models, including Indic-BERT, for hate speech detection in Devanagari Script Languages. (Paval et al., 2024) introduced a multimodal abuse detection system using Liquid Neural Networks for text and CNN for audio, achieving strong performance across 10 Indian languages.

3 Task and Dataset Description

For this shared task, a comprehensive dataset was provided to identify abusive language targeting women in Tamil and Malayalam social media text. The task identifies whether a given comment is abusive or non-abusive for better online content moderation. The dataset for this task consists of comments scraped from YouTube, covering explicit abuse, implicit bias, stereotypes, and coded language targeting women. Each comment is annotated with binary labels. The abusive comment detection dataset for Tamil was provided in the previous workshop (Priyadharshini et al., 2022), while the dataset for Tamil and Telugu was shared in the 2023 workshop (Priyadharshini et al., 2023).
Abusive: Content that conveys hateful, harassing, or derogatory language directed at women.
Non-Abusive: Content that does not contain hateful, harassing, or derogatory language.
Here, Table 1 reports the number of samples across the two categories.

Classes	Train	Valid	Test
Abusive	1,531	303	323
Non-Abusive	1,402	326	306
Total	2,933	629	629

Table 1: Statistical Distribution of Classes across Train, Validation, and Test Datasets.

The dataset is slightly imbalanced, with fewer non-abusive samples in the train dataset. The bar chart in Figure 1 represents the percentage of abusive and non-abusive comments.

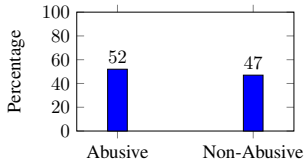


Figure 1: Statistics of training dataset.

4 Methodology

The section describes the methodology including data preparation, modeling, and evaluation phase. Malayalam-BERT was chosen based on its strong performance in prior NLP tasks, such as achieving the highest accuracy in fake news detection for Malayalam text classification (Tabassum et al., 2024). The schematic representation of our approach is depicted in Figure 2.

4.1 Preprocessing

In this stage, several steps were applied to clean and standardize the text data. First, we cleaned the text data by removing URLs, emojis, HTML tags, punctuation, and special characters. The whitespace was normalized, and all text was converted to lowercase for consistency.

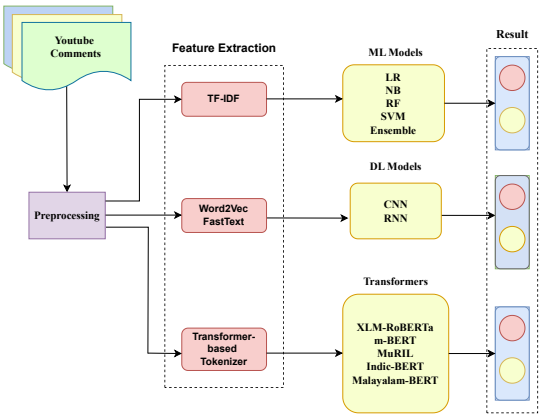


Figure 2: Abstract view of our methodology.

4.2 Feature Extraction

Feature extraction is conducted prior to training the models. For machine learning models, we employed Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988). For deep learning models, word embeddings were generated using Word2Vec (Mikolov et al., 2013) trained from scratch on our dataset, converting words into dense vector representations that capture semantic relationships. FastText embeddings were used for the RNN model, providing better word vectorization by considering subword information (Bojanowski et al., 2017).

4.3 Model Building

In our research, we explored several ML, DL, and transformer-based models.

4.3.1 ML models

We trained and evaluated algorithms using TF-IDF features. These include Logistic Regression (LR) (McFadden, 1972), Naïve Bayes (NB) (Maron, 1961), Support Vector Machines (SVM) (Liu et al., 2010), and Random Forest (RF) (Liaw et al., 2002). Additionally, we used a Voting Classifier ensemble combining LR, SVM, and RF to improve performance (Hossain et al., 2022).

4.3.2 DL models

In the case of the DL approach, we explored two architectures: a Convolutional Neural Network (CNN) (Chen et al., 2017) trained on Word2Vec embeddings and a Simple Recurrent Neural Network (SimpleRNN) model (Emon et al., 2019) that used FastText embeddings. The CNN was trained for 10 epochs and the SimpleRNN for 12 epochs, both with a batch size of 32 and fine-tuned using validation data.

4.3.3 Transformers

The transformer-based models, including MuRIL (Khanuja et al., 2021), Indic-BERT (Kakwani et al., 2020), XLM-R (Lample and Conneau, 2019) and m-BERT (Devlin et al., 2018) were used to identify abusive content in code-mixed Indic languages. Lastly, Malayalam-BERT, which has shown strong performance in fake news classification (Tripty et al., 2024), was also applied. These models are fine-tuned with transformer-specific tokenizers to handle multilingual text efficiently. Transformers outperform ML and DL models using attention mechanisms to capture context and dependencies.

5 Results & Discussion

Several machine learning, deep learning, and transformer models are experimented with using the given dataset. Naive Bayes, SVM, and an ensemble model performed best among ML models, while CNN and RNN underperformed. Transformers outperformed both, with Malayalam-BERT leading, followed by m-BERT and XLM-R, while MuRIL lagged. To optimize performance, we fine-tuned transformers using AdamW, training XLM-R for 15 epochs, m-BERT for 15 and 10 epochs, and Malayalam-BERT for 15 epochs, improving at 12 epochs in Table 2. After adjusting hyper-

Hyperparameters	XLM	m-BERT		Malayalam-BERT
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Learning rate	2e-06	3e-06	2e-06	2e-06	3e-06
Epochs	15	15	10	15	12
Batch size	32	16	32	32	16
Weight Decay	1e-04	1e-05	1e-06	1e-04	1e-04
Dropout	0.5	0.4	0.5	0.5	0.4

Table 2: Summary of tuned hyper-parameters.

parameters, Malayalam-BERT achieved the highest MF1 of 0.71 at 15 epochs. m-BERT performed best at 15 epochs, achieving a score of 0.67, while XLM-R reached a macro-F1 score of 0.64. MuRIL struggled with a score of 0.31. Indic-BERT scored 0.57 at 10 epochs, outperforming MuRIL but lagging behind m-BERT and Malayalam-BERT. The precision, recall, and macro-F1 scores for each model are summarized in Table 3.

5.1 Quantitative Discussion

The results highlight the effectiveness of Malayalam-BERT in detecting abusive Malay-

Classifier	P	R	MF1
LR	0.64	0.64	0.64
NB	0.65	0.65	0.65
RF	0.61	0.61	0.61
SVM	0.65	0.65	0.65
Ensemble	0.65	0.65	0.65
CNN	0.49	0.50	0.46
RNN	0.45	0.46	0.43
XLM	0.67	0.65	0.64
m-BERT	0.68	0.67	0.67
MuRIL	0.50	0.22	0.31
Indic-BERT	0.59	0.58	0.57
Malayalam-BERT	**0.71**	**0.71**	**0.71**

Table 3: Performance of explored models.

alam text targeting women. Malayalam-BERT outperformed other transformer models like m-BERT and XLM-R due to its targeted training in Dravidian languages, allowing it to better understand the linguistic nuances of Malayalam. While m-BERT and XLM-R are multilingual models, their broader training scope leads to less precise detection of abusive language in Malayalam. Indic-BERT performed moderately better than MuRIL, which showed much lower scores. Although there is a slight class imbalance, we addressed this by applying class weights during training. The confusion matrix is shown in Figure 3. The model correctly classifies 239

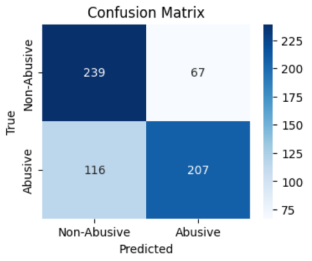


Figure 3: Confusion matrix of our best performing model.

Non-Abusive and 207 Abusive instances, but misclassifies 67 Non-Abusive instances as Abusive and 116 Abusive instances as Non-Abusive. These misclassifications may be due to class imbalance, where the model is biased toward the majority class, and limited data diversity.

5.2 Qualitative Discussion

Table 4 highlights both correctly classified and misclassified cases. Among the misclassified cases:

False Positives:

- **Sample 1** ("I loved you so much, now it's like full-blown acting.") expresses emotional disappointment but isn't abusive. The misclassification suggests the model struggles with emotionally charged non-abusive language.
- **Sample 3** ("No more lying on the ground and rolling around... We are not stupid...") uses negative words like 'stupid,' but not in an abusive way.

False Negatives:

- **Sample 7** ("You don't deserve 50 lakhs. There were people out there who deserved it more than you.") questions someone's worthiness without explicit offensive language, which the model fails to recognize as abuse.
- **Sample 8** ("If it was meant seriously, it would have been a great comedy.") is sarcastic ridicule that the model misses due to lack of explicit offensive words.

Test Sample	Actual	Predicted
Sample 1: നിങ്ങളെ ഒരുപാട് ഇഷ്ട്ടം ആയിരുന്നു ഇപ്പോ ഫുൾ അഭിനയം പോലെ ആണ് (I loved you so much, now it's like full-blown acting.)	0	1
Sample 2: ഇതൊക്കെ ആരാ എഴുതി തരുന്നത് കണ്ടാ പറയില്ലട്ടാ ബയങ്കര ഓർജിനാലിട്ടി (I can't tell you who wrote all this, Bayankara Orginalitti.)	1	1
Sample 3: ഇനി വീണസ്ഥലത്ത് കിടന്ന് ഉരുളണ്ട..... ഞങ്ങൾ മണ്ടന്മാരല്ല...(No more lying on the ground and rolling around... We are not stupid...)	0	1
Sample 4: "കൈ വിട്ട ആയുധവും, വാ വിട്ട വാക്കും തിരിച്ചെടുക്കാൻ കഴിയില്ല." ("A weapon once lost cannot be taken back, nor can a word once lost.")	0	0
Sample 5: ഈ പ്രശ്നത്തിൽ ഇടപെടാൻ ഈ സൂരജ് മൈരൻ ആരാണ് (Who is this Suraj Mairan to interfere in this issue?)	0	0
Sample 6: ഒരുപാട് ഇഷ്ടപ്പെട്ടിരിന്നു നിങ്ങളുടെ കോംബോ പക്ഷേ നിങ്ങളുടെ ഇതിനു മുൻപുള്ള വീഡിയോയിൽ നല്ല രീതിയിൽ പറഞ്ഞു അവസാനിപ്പികാമായിരുന്നു. (I really liked your combo, but you could have ended it in a better way in your previous video.)	0	0
Sample 7: നീ 50ലക്ഷത്തിനു അർഹയല്ല. നിന്നെക്കാൾ അർഹതയുള്ളവർ അവിടെ അവിടെ ഉണ്ടായിരുന്നു (You don't deserve 50 lakhs. There were people out there who deserved it more than you.)	1	0
Sample 8: സീരിയസ് ആയിട്ട് പറഞ്ഞതാണെങ്കിൽ വൻ കോമഡി ആയിട്ടുണ്ട് (If it was meant seriously, it would have been a great comedy.)	1	0

Figure 4: Examples of the Malayalam-BERT model's anticipated outputs with English translations.

These misclassifications indicate the model's struggle with indirect abuse and sarcasm.

6 Conclusion

Our study highlights the effectiveness of Malayalam-BERT in detecting abusive language targeting women on Malayalam social media, outperforming traditional ML and DL models with an F1 score of 0.71. In future work, we intend to improve accuracy and F1 score through advanced feature extraction and augmentation. While focused on Malayalam, our methodology can be adapted to other low-resource languages using models like m-BERT, XLM-R, IndicBERT, or MuRIL. Furthermore, we will investigate the integration of multimodal approaches, incorporating textual, visual, and audio cues to improve abusive content detection particularly for social media.

Limitations

Our model's performance is affected by certain constraints. While deep learning models like CNN and RNN underperformed compared to transformer-based models like Malayalam-BERT, this highlights their inefficiency for complex text classification tasks. We trained embeddings from scratch on this small Malayalam dataset, which may result in sparse and ineffective representations, whereas pre-trained FastText or Word2Vec are trained on massive corpora and capture richer semantic and syntactic relationships. A key limitation is the handling of Out-of-Vocabulary (OOV) words, particularly in informal social media text. The tokenizer may struggle with Malayalam's rich morphology and misspelled or unique words, impacting performance. Subword tokenization or domain-specific vocabulary could mitigate this issue.

References

Raneem Alharthi, Rajwa Alharthi, Ravi Shekhar, and Arkaitz Zubiaga. 2023. Target-oriented investigation of online abusive attacks: A dataset and analysis.

Premjith B, Jyothish G, Sowmya V, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, Bharathi B, Saranya Rajiakodi, Rahul Ponnusamy, Jayanth Mohan, and Mekapati Reddy. 2024. Findings of the shared task on multimodal social media data analysis in Dravidian languages (MSMDA-DL)@DravidianLangTech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61, St. Julian's, Malta. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Hao Chen, Susan Mckeever, and Sarah Jane Delany. 2017. Abusive text detection using neural networks. In *Irish Conference on Artificial Intelligence and Cognitive Science*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Estiak Ahmed Emon, Shihab Rahman, Joti Banarjee, Amit Kumar Das, and Tanni Mittra. 2019. A deep learning approach to detect abusive bengali text. In *2019 7th International Conference on Smart Computing Communications (ICSCC)*, pages 1–5.

Alamgir Hossain, Mahathir Bishal, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. COMBATANT@TamilNLP-ACL2022: Fine-grained categorization of abusive comments using logistic regression. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 221–228, Dublin, Ireland. Association for Computational Linguistics.

Horacio Jarquın-Vasquez, Hugo Jair Escalante, Manuel Montes-y Gomez, and Fabio A. Gonzalez. 2024. Gha: a gated hierarchical attention mechanism for the detection of abusive language in social media. *IEEE Transactions on Affective Computing*, pages 1–14.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vishnu Subramanian, and Partha Pratim Talukdar. 2021. Muril: Multilingual representations for indian languages. *ArXiv*, abs/2103.10730.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *ArXiv*, abs/1901.07291.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

Zhijie Liu, Xueqiang Lv, Kun Liu, and Shuicai Shi. 2010. Study on svm compared with the other text classification methods. In *2010 Second International Workshop on Education Technology and Computer Science*, volume 1, pages 219–222.

Hema M, Anza Prem, Rajalakshmi Sivanaiah, and Angel Deborah S. 2023. Athena@DravidianLangTech: Abusive comment detection in code-mixed languages using machine learning techniques. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 147–151, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

M. E. Maron. 1961. Automatic indexing: An experimental inquiry. *J. ACM*, 8:404–417.

Daniel McFadden. 1972. Conditional logit analysis of qualitative choice behavior.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Vasanth Palanikumar, Sean Benhur, Adeep Hande, and Bharathi Raja Chakravarthi. 2022. DE-ABUSE@TamilNLP-ACL 2022: Transliteration as data augmentation for abuse detection in Tamil. In *Proceedings of the Second Workshop on Speech and*

Language Technologies for Dravidian Languages, pages 33–38, Dublin, Ireland. Association for Computational Linguistics.

Ks Paval, Vishnu Radhakrishnan, Km Krishnan, G Jyothish Lal, and B Premjith. 2024. Multimodal fusion for abusive speech detection using liquid neural networks and convolution neural network. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7.

Sara Jerin Prithila, Fariha Hasan Tonima, Tahsina Tajrim Oishi, Md. Nazrul Islam, Ehsanur Rahman Rhythm, Adib Muhammad Amit, and Annajiat Alim Rasel. 2023. Detecting derogatory comments on women using transformer-based models. In *2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pages 278–284.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.

Deepawali Sharma, Vedika Gupta, and Vivek Kumar Singh. 2024. 11 - abusive comment detection in tamil using deep learning. In D. Jude Hemanth, editor, *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications*, pages 207–226. Morgan Kaufmann.

Rajalakshmi Sivanaiah, Rajasekar S, Srilakshmisai K, Angel Deborah S, and Mirnalinee ThankaNadar. 2023. Avalanche at DravidianLangTech: Abusive comment detection in code mixed data using machine learning techniques with under sampling. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 166–170, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_Punctuators@DravidianLangTech-EACL2024: Transformer-based approach for detection and classification of fake news in Malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186, St. Julian's, Malta. Association for Computational Linguistics.

Farjana Alam Tofa, Lorin Tasnim Zeba, Md Osama, and Ashim Dey. 2025. CUET_INSights@NLU of Devanagari script languages 2025: Leveraging transformer-based models for target identification in hate speech. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 267–272, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Zannatul Tripty, Md. Nafis, Antu Chowdhury, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUETSentimentSillies@DravidianLangTech EACL2024: Transformer-based approach for detecting and categorizing fake news in Malayalam language.

Suppawong Tuarob, Manisa Satravisut, Pochara Sangtunchai, Sakunrat Nunthavanich, and Thanapon Noraset. 2023. Falcon: Detecting and classifying abusive language in social networks using context features and unlabeled data. *Information Processing Management*, 60(4):103381.

Advaitha Vetagiri, Gyandeep Kalita, Eisha Halder, Chetna Taparia, Partha Pakray, and Riyanka Manna. 2024. Breaking the silence detecting and mitigating gendered abuse in hindi, tamil, and indian english online spaces. *Preprint*, arXiv:2404.02013.

Mohammad Zia Ur Rehman, Somya Mehta, Kuldeep Singh, Kunal Kaushik, and Nagendra Kumar. 2023. User-aware multilingual abusive content detection in social media. *Information Processing Management*, 60(5):103450.

SemanticCuetSync@DravidianLangTech 2025: Multimodal Fusion for Hate Speech Detection - A Transformer Based Approach with Cross-Modal Attention

Md. Sajjad Hossain, Symom Hossain Shohan, Ashraful Islam Paran , Jawad Hossain and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904031, u1904048, u1904029, u1704039}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

The rise of social media has significantly facilitated the rapid spread of hate speech. Detecting hate speech for content moderation is challenging, especially in low-resource languages (LRLs) like Telugu. Although some progress has been noticed in hate speech detection in Telegu concerning unimodal (text or image) in recent years, there is a lack of research on hate speech detection based on multimodal content detection (specifically using audio and text). In this regard, DravidianLangTech has arranged a shared task to address this challenge. This work explores three machine learning (ML), three deep learning (DL), and seven transformer-based models that integrate text and audio modalities using cross-modal attention for hate speech detection. The evaluation results demonstrate that mBERT achieved the highest F-1 score of 49.68% using text. However, the proposed multimodal attention-based approach with Whisper-small+TeluguBERT-3 achieves an F-1 score of 43 68%, which helps us achieve a rank of 3rd in the shared task competition.

1 Introduction

Social media platforms have emerged as the focal point for information sharing in the rapidly evolving digital world, where individuals interact and communicate. On the one hand, increased connectivity and easier idea sharing have resulted from increased online activities, and it has also accelerated the spread of hate speech and other forms of internet harassment. Hate speech refers to communication, including speaking, writing, and symbolic expressions, that spreads hatred, slander, discrimination, and violence. It may be aimed at a specific person or group based on traits including race, color, ethnicity, religion, gender, sexual orientation, caste, country, or socioeconomic class (Nockleby, 1994; Keipi et al., 2016; Benikova et al., 2018). Due to the large volume of data, manually monitoring and identifying hate speech is impractical. That is why manual moderation is impractical. Thus an automatic system for hate speech detection is essential for real-time detection of harmful content and creating a safer online space.

The subjective and context-dependent nature of hate speech makes detecting hate speech a complex problem. As the meaning of specific phrases varies across cultures and social and situational factors, it becomes more challenging to understand the context. Sometimes, it is tough to distinguish between hate speech and legitimate expressions like satire or criticism. This problem often requires a nuanced understanding of the language. Also, certain words or slang in social media are uncommon in daily conversation, making it difficult to identify as hate speech. Various research has been conducted in the Natural Language Processing (NLP) domain to detect hate speech. Most previous work concentrated on a single domain like text or audio (Alkomah and Ma, 2022; Imbwaga et al., 2024). The multimodal aspects of the problem make it even more difficult. We proposed a cross-modal attention-based approach to fuse text and audio in this shared task on Multimodal Hate Speech Detection in Dravidian languages (Premjith et al., 2024a,b). The main contributions of this work are:

- Proposed a cross-modal attention-based approach to fuse two modalities for hate speech detection in Telugu.
- Investigated several transformers and DL models for hate speech detection in Telugu exploiting textual and audio features.

2 Related Work

Many studies have been conducted in recent years to identify hate speech. Sreelakshmi et al., 2024 presented a mix of multilingual transformer-based

embedding models with ML classifiers to detect hate speech and foul language in CodeMix Dravidian languages. After examining models such as MuRIL, BERT, and XLM, they discovered that MuRIL, combined with an SVM classifier, achieved the best performance across Kannada-English, Malayalam-English, and Tamil-English datasets, with accuracies up to 96%. Their study also featured a cost-sensitive learning strategy to address class imbalance, as well as a novel annotated Malayalam-English CodeMix dataset. Hakim et al., 2024 presented a combination of transformer and deep learning models to identify hate speech in Indonesian tweets. Combining IndoBERTweet, BiLSTM, and CNN resulted in an F-1 score of 85.06%

Talking about multiple modalities, Arya et al., 2024 have identified hate speech in memes using the Contrastive Language-Image Pre-Training (CLIP) model with prompt engineering. They have used the Facebook Hateful Meme dataset (Kiela et al., 2020), which contains two modalities (Text and Image). Their finetuned CLIP model scored F-1 score of 90.12%. Mandal et al., 2024 also proposed a technique for identifying hate speech using transformers. Their dataset also contains two modality, but this time, audio and text (English). They have used a new fusion technique called Attentive Fusion, which helped their model to get F-1 score of 92.70%. Similarly, Imbwaga et al., 2024 offered numerous machine learning-based approaches to identify hate speech in English and Kiswahili from audio. The Extreme Gradient Boosting Model achieved the highest F-1 score (96.10%) in Kiswahili, whereas Random Forest achieved the highest F-1 score (90.00%) in English.

There has been a lack of research on identifying hate speech in Telugu using audio and text. This work developed a multimodal framework leveraging transformers to bridge this gap.

3 Dataset and Task Description

This task (Lal G et al., 2025) mainly focused on creating models that accurately detect hate speech in Telugu speech and texts. This work used a multimodal hate speech dataset created by Anilkumar et al., 2024. The dataset includes five hate speech classes: Gender (G), Political (P), Religious (R), Personal Defamation (PD), and Non-hate (NH). The definition (Sharif et al., 2022) of the classes are illustrated in the following:

- **Gender (G):** Use offensive references to body parts, sexual orientation, sexuality, or other pornographic material to harm a person or group.
- **Political (P):** Criticize political ideologies, provoke party supporters, or stir people against the government and police enforcement.
- **Religious (R):** Provoke violence by insulting a religion, religious group, or religious beliefs (Catholic, Hindu, Jewish, or Islamic, among others).
- **Personal Defamation (PD):** Act of making false statements about an individual that harm their reputation.
- **Non-hate (NH):** Do not make any rude comments or convey any hostile intent to hurt other people mentally or physically.

Modality	**Train**	**Test**	**Total**
Text	556	50	606
Audio	551	50	601
Total	1107	100	1207
T_W **(Text)**	13170	1064	14234
T_{UW} **(Text)**	6598	696	7294
T_{avg} **(Text)**	23	21	–
A_{avg} **(Audio)**	1055	918	–
D_{avg} **(Audio)**	12	10	–

Table 1: Dataset statistics for Task-3. The symbols T_W and T_{UW} denote the total and unique words in the text, whereas A_{avg} indicates the average audio size in KB and D_{avg} indicates the average duration of audio in seconds.

The dataset comprises 556 texts and 551 audio in the training set and 50 texts and 50 audio in the test set. Task-3 concerns multimodal hate speech detection in Telugu. Table 1 shows the distribution of dataset into train, and test sets. The source code is publicly available at https://github.com/ashrafulparan2/SemanticCuetSync-DravidianLangTech-2025.

4 System Overview

This work exploited several transformer-based models to address task 3. Textual and audio features train the models and fuse the outputs with cross-modal attention. Figure 1 illustrates the

schematic configuration of the proposed multimodal hate speech detection solution.

4.1 Feature Extraction

The feature extraction involves two independent processes for text and audio modalities.

4.1.1 Text

We have investigated 8 transformer-based models and 1 DL model for textual feature extraction.

- **BERT:** This (Devlin, 2018) transformer-based model was pre-trained and self-supervised on a large corpus of English data. The training process used raw texts and was conducted without human labeling, employing two objectives: masked language modeling (MLM) and next-sentence prediction. As a result, this model has developed a robust understanding of the language's internal representation. In this task, this model was employed for feature extraction.

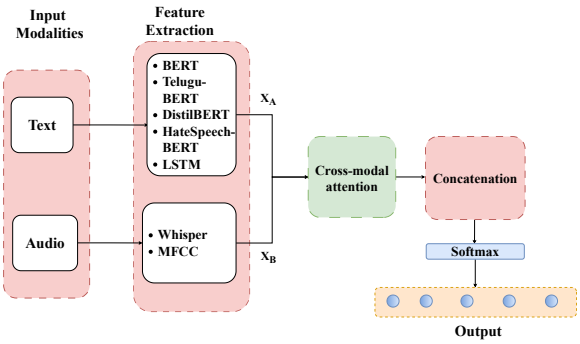


Figure 1: Schematic process for hate speech detection.

- **TeluguBERT (TBERT):** We used five versions of TeluguBERT (Joshi, 2022) for textual feature extraction in this task. This BERT model is trained on a publicly available Telugu monolingual dataset. The extensive training enables the model to capture the rich linguistic nuances, syntax, and semantic patterns unique to the Telugu language, which can be very useful for textual feature extraction.

- **DistilBERT (dBERT):** This (Sanh, 2019) is a smaller, faster, cheaper, and lighter version of the BERT model. A notable characteristic of this model is that it has 40% fewer parameters than the BERT models. As the number of parameters is lower, it is 60% faster. Most importantly, it maintains 95% of BERT's performance as measured on the GLUE language understanding benchmark. We used this model for textual feature extraction in this task.

- **HateSpeechBERT (HS-BERT):** This is a pre-trained BERT-based model specially fine-tuned for detecting abusive speech in Bengali, Devanagari Hindi, code-mixed Hindi, code-mixed Kannada, code-mixed Malayalam, Marathi, code-mixed Tamil, Urdu, and English. We used this model for textual feature extraction.

Table 2 illustrates the hyperparameters used in transformer-based models. The hyperparameters were tuned manually based on empirical observations and iterative experimentation.

Models	LR	WD	WS	EP
Unimodal (text)	5e-5	0.30	50	10
Unimodal (Audio)	3e-5	0.01	0	5
Bimodal	1e-5	0.00	0	10

Table 2: Hyperparameters for transformer-based models.

4.1.2 Audio Features

- **Whisper:** Whisper (Radford et al., 2023) is a state-of-the-art pre-trained model developed for automatic speech recognition (ASR). It is also trained for speech translation. This model is trained on approximately 680k hours of labeled data. This vast training corpus enables Whisper to demonstrate a strong ability to generalize to many datasets and domains. We used this model for auditory feature extraction because of its robust performance and multilingual capabilities.

- **MFCC:** Mel-frequency Cepstral Coefficients (MFCC) is another popular auditory feature extractor used in this task for detecting hate speech. It is designed to mimic the way humans perceive sound and speech. It analyzes the power spectrum of the audio signal and maps it to the Mel scale.

4.2 Cross-modal Attention

After the feature extraction steps, we used a cross-modal attention (Ye et al., 2019) mechanism between the audio-text pair. Cross-modal attention can be represented by the Eqs. (1)-(5).

1. Query, Key, and Value Projections:

$$Q = Z_A W_Q \quad (1)$$

$$K = Z_B W_K \quad (2)$$

$$V = Z_B W_V \quad (3)$$

2. Scaled Dot-Product Attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (4)$$

3. Concatenation:

$$\text{Output} = \text{Concat}(\text{Attention}(Q, K, V), \dots) \quad (5)$$

The equation of concatenation:

$$\alpha_{\text{concat}} = [\alpha_1; \alpha_2; \cdots; \alpha_n] \quad (6)$$

Here, Z_A represents features from modality A, and Z_B represents features from modality B.

4.3 Fusion

In this step, we concatenated the output from the cross-modal attention layers. This is used to produce the final output. Equation 7, the early fusion approach, concatenates audio and text features.

$$F = [F_{\text{audio}} \oplus F_{\text{text}}] \quad (7)$$

Here, F is the fused feature representation, F_{audio} represents the feature vector extracted from the audio modality, F_{text} represents the feature vector extracted from the text modality, and $\oplus$ denotes the concatenation operation.

5 Results and Analysis

Table 3 demonstrates the evaluation results of unimodal and bimodal models on the test set.

Among unimodal (Text) models, mBERT surpasses all others with the highest F-1 score of 49.68%. dBERT scores the lowest, with an F-1 score of 17.46%. We analyzed numerous TBERT versions, and TBERT-5 had the highest F-1 score of 38.10%. Among unimodal (Audio) models, Hubert surpasses all others with an F-1 score of 22.94%.

For Bimodal (Audio+Text), we have explored several transformer-based models with early fusion. Whisper-small and TBERT-3 with early fusion outperform all other models with an F-1 score of 43.68%. However, for TBERT versions 4 and 5, the F-1 score decreases gradually. Whisper-small with HS-BERT results in the lowest F-1 score of 28.12%. Appendix D illustrates the detailed error analysis of the best-performed models (mBERT and Whisper-small+TBERT-3).

Unimodal (Text)				
Classifier	**Pr(%)**	**Re(%)**	**F1(%)**	**Ac(%)**
SVM	68.25	32.63	31.60	48.65
Random Forest	28.80	37.20	32.18	52.25
Logistic Regression	68.70	37.47	37.79	53.15
CNN	46.91	38.80	33.58	56.76
CNN + LSTM	30.07	38.00	31.01	51.35
CNN + BiLSTM	36.52	43.20	39.12	58.56
TBERT-1	36.42	40.00	37.20	40.00
TBERT-2	40.00	40.00	37.98	40.00
TBERT-3	34.00	34.38	34.00	33.57
TBERT-4	32.00	47.28	32.00	29.96
TBERT-5	40.61	38.00	38.10	38.00
dBERT	12.52	30.00	17.46	30.00
mBERT	50.94	50.00	**49.68**	50.00
Unimodal (Audio)				
Classifier	**Pr(%)**	**Re(%)**	**F1(%)**	**Ac(%)**
Whisper-small	14.86	20.00	17.00	20.00
Hubert	17.13	38.00	**22.94**	38.00
Wav2vec2	7.74	20.00	10.34	20.00
Bimodal				
Classifier	**Pr(%)**	**Re(%)**	**F1(%)**	**Ac(%)**
Whisper-small + BERT	34.88	38.00	33.58	38.00
Whisper-small + TBERT-1	42.69	38.00	32.78	38.00
Whisper-small + TBERT-2	40.00	40.00	37.98	40.00
Whisper-small + TBERT-3	43.44	46.00	**43.68**	46.00
Whisper-small + TBERT-4	21.87	36.00	27.05	36.00
Whisper-small + TBERT-5	31.28	46.00	35.21	46.00
Whisper-small + dBERT	32.90	46.00	36.47	46.00
Whisper-small + HS-BERT	40.00	22.18	28.12	40.00
MFCC + LSTM	13.63	18.00	9.25	18.00

Table 3: Performance of the employed models for the tasks.

6 Error Analysis

We have analyzed the proposed model's performance to illustrate a quantitative and qualitative error analysis.

Quantitative Analysis

Figure 2 depicts the confusion matrix for the test set, categorizing speeches into their appropriate classes. The findings suggest that 23 out of 50 speeches were properly predicted. Among the five categories, "Personal Defamation" was the most precisely identified, while "Gender" and "Political" hate speeches were only correctly identified once each. Overall performance was unsatisfactory, owing to the dataset's limited size.

Qualitative Analysis

Figure 3 displays predicted outputs and their corresponding true labels for some randomly selected samples, demonstrating the proposed model's performance. The model frequently struggles to appropriately understand the intent underlying the tone of a speech. The same speech may convey multiple meanings, depending on the tone in which it is delivered. For example, in the second case, the model failed to catch the subtle tone of the speech,

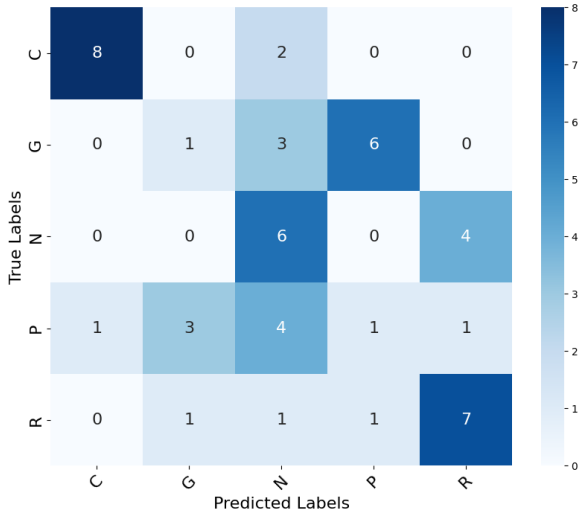


Figure 2: Confusion matrix of the best performing model.

resulting in a misclassification. This shortcoming is mostly due to the model is trained on a very limited dataset, which limits its capacity to accurately recognize hate speech.

Transcript	AL	PL
ఎవరు మాత్రం SC కులంలో పుట్టాలని కోరుకుంటారు (Who wants to be born in SC caste)	Religious	Religious
ఎస్సీలుగా పుట్టాలని ఎవరు కోరుకుంటారు (Who wants to be born as SCs?)	Religious	Gender
నీ బండారం బయటపెట్టి పొలం మధ్య నిలబెట్టి గుడ్డలు ఊడదీస్తా (You will take out your barn and stand it in the middle of the field and blow the rags)	Personal Defamation	Personal Defamation
హిందువులెవ్వరు ముస్లింలకు వ్యతిరేకం కాదని కూడా ఆయన తన అభిప్రాయంగా చెప్పారు (He also said in his opinion that no Hindu is against Muslims	Non-Hate	Non-Hate

Figure 3: Few randomly selected samples with actual (AL) and predicted labels (PL).

7 Conclusion

This study investigated several transformers and DL techniques in both audio and text modality with cross-modal attention for detecting hate speech in Telugu. Among unimodal (Audio) models, Hubert surpasses all others with an F-1 score of 22.94%. Among various bimodal (audio + text) combinations, Whisper-small + TeluguBERT-3 achieved the highest F1 score of 43.68%. However, we found that mBERT achieves a higher F1 score of 49.68% using text only. This study demonstrates that the textual unimodal approach gives us a superior performance. Further improvements can be made by increasing the dataset and using other multimodal models. Besides, exploring various LLMs may improve results for detecting Telugu hate speech.

Limitations

The current implementations possess some weaknesses, such as (i) The dataset is limited in size, so the model suffers from generalization issues, and (ii) the noise and recording quality of audio data affect performance.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Abhishek Anilkumar, Jyothish Lal G, B Premjith, and Bharathi Raja Chakravarthi. 2024. Dravlangguard: A multimodal approach for hate speech detection in dravidian social media. In *Speech and Language Technologies for Low-Resource Languages (SPELLL)*, Communications in Computer and Information Science.

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2018. What does this imply? examining the impact of implicitness on the perception of hate speech. In *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings 27*, pages 171–179. Springer.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Atalla Naufal Hakim, Yuliant Sibaroni, and Sri Suryani Prasetyowati. 2024. Detection of hate-speech text on indonesian twitter social media using indobertweet-bilstm-cnn. In *2024 12th International Conference on Information and Communication Technology (ICoICT)*, pages 374–381. IEEE.

Joan L Imbwaga, Nagatatna B Chittaragi, and Shashidhar G Koolagudi. 2024. Automatic hate speech detection in audio using machine learning algorithms. *International Journal of Speech Technology*, 27(2):447–469.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Teo Keipi, Matti Näsi, Atte Oksanen, and Pekka Räsänen. 2016. *Online hate and harmful content: Cross-national perspectives*. Taylor & Francis.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Atanu Mandal, Gargi Roy, Amit Barman, Indranil Dutta, and Sudip Kumar Naskar. 2024. Attentive fusion: A transformer-based approach to multimodal hate speech detection. *arXiv preprint arXiv:2401.10653*.

John T Nockleby. 1994. Hate speech in context: The case of verbal threats. *Buff. L. Rev.*, 42:653.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2022. M-bad: A multilabel dataset for detecting aggressive texts and their targets. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 75–85.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511.

A Class-wise Distribution of Dataset

Figure A.1 presents the class-wise distribution of the dataset, illustrating the frequency of samples across five distinct categories: Non-hate, Personal Defamation, Gender, Religious, and Political. The Non-hate category constitutes the largest proportion, with 198 instances, followed by Personal Defamation (122), Gender (101), Religious (72), and Political (58). This distribution highlights a class imbalance, with Non-hate being the dominant class, which may influence model training and performance.

Figure A.2 illustrates a few examples of the input and output of the dataset.

B System Requirements

This study was developed using Python 3 (version 3.10.12) and Python-based libraries from the PyTorch 2 framework to implement transformers, including BERT, TBERT, dBERT, and Whisper-small. The implementation required 29GB of RAM, 16GB of VRAM, and 73.1GB of storage space. We utilized an NVIDIA Tesla P100 GPU on Kaggle. For data analysis and preprocessing, we employed pandas (2.1.4) and numpy (1.24.3). For unimodal, ML models were built using scikit-learn (1.2.2), while DL models were trained with Keras

(2.13.1) and TensorFlow (2.13.0). Additionally, PyTorch (2.0.0) and transformers (4.36.2) implement transformer-based bimodal models.

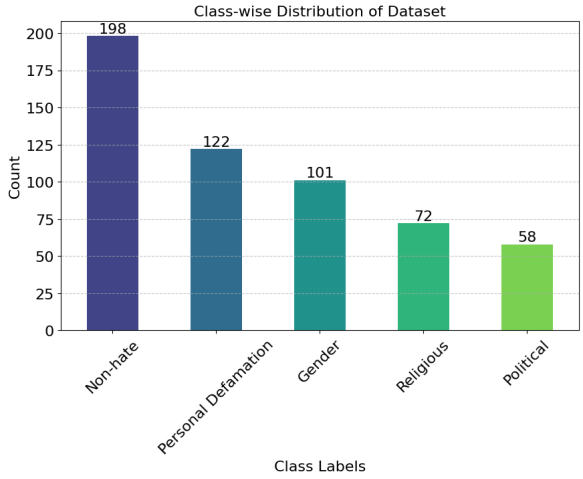


Figure A.1: Class-wise Distribution of the training dataset.

Transcript	Label
ఏసుప్రభు నిజమైన దేవుడు అని చెప్తాకే వచ్చాను నేను. ఛీ తీయ్ బండి తీయ్ (I have come to tell you that Jesus is the true God. Chee Tee Bandi Tee)	Religious
వైయస్సార్ కాంగ్రెస్ అనగానే ఆగుమాటిక్ గ అవినీతి, అది రెండు పార్టీలు మధ్య ఉన్న వేత్తయసం. (YSR Congress is an autocratic corruption, it is a conflict between two parties.	Political
నా మీద బతికి ఉన్నా గోజ్జ లంగాకొడుకులారా (Live on me, you bastards)	Gender
ఒక మనిషిని కదిలించే శక్తి సహితాయ్నకి మతమే ఉంటుంది అక్షరాన్ని మతమే ఉంటుంది (Religion is the power that moves a man, religion is the letter)	Non-Hate

Figure A.2: Task-3 sample with Transcript and label.

CUET_Novice@DravidianLangTech 2025: A Bi-GRU Approach for Multiclass Political Sentiment Analysis of Tamil Twitter (X) Comments

Arupa Barua, Md Osama and **Ashim Dey**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u2004089, u1804039}@student.cuet.ac.bd, ashim@cuet.ac.bd

Abstract

Multilingual political sentiment analysis faces challenges in capturing subtle variations, especially in complex and low-resourced languages. Identifying sentiments correctly is crucial to understanding public discourse. A shared task on Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments, organized by DravidianLangTech@NAACL 2025, provided an opportunity to tackle these challenges. For this task, we implemented two data augmentation techniques, which are synonym replacement and back translation, and then explored various machine learning (ML) algorithms. We experimented with deep learning (DL) models including GRU, BiLSTM, BiGRU, hybrid CNN-GRU and CNN-BiLSTM to capture the semantic meanings more efficiently using FastText and CBOW embedding. The Bidirectional Gated Recurrent Unit (BiGRU) achieved the best macro-F1 (MF1) score of 0.33, securing the 17th position in the shared task. These findings underscore the challenges of political sentiment analysis in low-resource languages and the need for advanced language-specific models for improved classification.

1 Introduction

Political sentiments are the views and feelings expressed by individuals or groups about political issues. Classifying political sentiments is crucial to understand public perspectives and addressing a variety of points of view. In multilingual contexts, sentiment analysis in Tamil is especially crucial due to the linguistic and cultural nuances that shape sentiment expression. The shared task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments organized by DravidianLangTech@NAACL 2025 aimed to address this challenge by identifying the types of political sentiments into seven classes: Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral and None of the above. Their workshop paper (Hegde et al., 2023) provided us an opportunity to engage with these challenges in processing South Asian languages and to leverage our work on political multiclass sentiment analysis.

In our participation, we focus on addressing the challenges of political sentiment classification through two primary contributions:

- We implement data augmentation in two steps to bring more diversity and balance to the training data.
- We leverage different machine learning and deep learning approaches to better capture contextual nuances and improve the overall accuracy of political sentiment classification.

Our code, developed for this shared task can be accessed at `https://github.com/ArupaBarua/DravidianLangTech-NAACL-Sentiment`.

2 Related Work

The complexity of understanding and categorizing political sentiments has driven extensive research employing various languages, datasets, and methodologies. Research has been done to advance sentiment analysis in under-resourced code-mixed languages (Sambath Kumar et al., 2024). Different machine learning models have been employed to classify sentiments in highly under-resourced, code-mixed languages like Tulu and Tamil (Shetty, 2023), (Shanmugavadivel et al., 2022a), (Kanta and Sidorov, 2023), (Ponnusamy et al., 2023), (Thavareesan and Mahesan, 2021). A grid search approach has been explored for analyzing sentiments in code-mixed Tamil and Telulu (B et al., 2024). ML models like SVMs and VSMs have been explored to analyze multiclass sentiments on short texts (K. Suresh Kumar and Moshayedi, 2024). Due to the inefficiency of machine learning models in extracting contextual meanings, their works lack the ability to fully capture nuanced expressions and complex sentiment patterns. Deep learning

models, CNN, and LSTM are particularly significant in this case as they excel at capturing complex patterns and contextual nuances in code-mixed languages (Rajasekar and Geetha, 2023), (Nithya et al., 2022), (Mandalam and Sharma, 2021). For multiclass sentiment analysis, Bidirectional Recurrent Neural Network (BiRNN) and its variations like BiLSTM have also been experimented (Krosuri and Aravapalli, 2024), (Roy and Kumar, 2021). To enhance the performance of ML and DL algorithms, different hybrid models have also been explored for Tamil sentiment analysis (Ramesh Babu, 2022), (Gandhi et al., 2021), (Shanmugavadivel et al., 2022b). Recently, transformer-based models like m-BERT, MiniLM, and Indic-BERT have been applied to hate speech detection, demonstrating improved contextual understanding and classification accuracy (Tofa et al., 2025). Multilingual transformers have been explored for multiclass sentiment analysis in code-mixed data, effectively capturing contextual nuances in low-resource languages (Nazir et al., 2025).

3 Task and Dataset Description

This shared task focuses on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments. The task organizers provided a dataset comprising X (Twitter) comments in Tamil language, which are annotated with seven categories (Chakravarthi et al., 2025). The objective is to classify these sentiments into seven labels, which are as follows:

- **Substantiated** – Sentiment backed by evidence, reference or logical reasoning.
- **Sarcastic** – Sentiment expressed in a mocking or ironic tone.
- **Opinionated** – Sentiment based on personal beliefs or viewpoints.
- **Positive** – Sentiment expressing approval or good feeling towards a political entity.
- **Negative** – Sentiment expressing criticism.
- **Neutral** – Sentiment that is impartial or does not express a strong emotion.
- **None of the above** – Sentiment that does not fit into any of the specified categories.

The distribution of the political sentiment classes across the training, development and test datasets is shown in Table 1. The training dataset exhibits a noticeable class imbalance, with the "Opinionated" category having the highest representation, significantly outnumbering other classes. In contrast, categories like "None of the above" and "Substantiated" have relatively fewer samples, which is illustrated in Figure 1.

Table 1: Class distribution across datasets.

Class	Train	Dev	Test
Opinionated	1361	153	171
Sarcastic	790	115	106
Neutral	637	84	70
Positive	575	69	75
Substantiated	412	52	51
Negative	406	51	46
None of the above	171	20	25

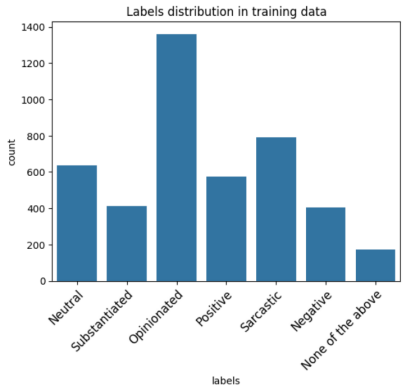


Figure 1: Class distribution in the train set

4 Methodology

The methods and strategies applied to predict the classes of political sentiments are discussed in this section. Through thorough analysis, we propose a Bidirectional Gated Recurrent Unit (BiGRU) network to estimate the multiclasses of political sentiments. Figure 2 provides a visualization of our methodology, outlining the key steps involved.

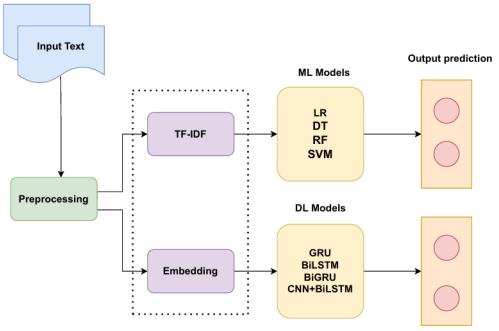


Figure 2: An abstract view of our methodology.

4.1 Preprocessing

In this step, several techniques were applied to refine the X comments. We cleaned the text by removing URLs, emojis, HTML tags, punctuation and special characters, normalized white spaces,

and converted all text to lowercase. To address the class imbalance, we implemented data augmentation on the training data in two steps. First, we applied synonym replacement with the FastText Tamil model. For each word, we retrieved its nearest synonym from the pre-trained model and replaced it accordingly to enhance diversity. Then, to improve predictions for minority classes, we applied back-translation to the underrepresented categories 'None of the above', 'Negative', and 'Substantiated' using the mBART model (Tang et al., 2020) and then implemented RandomOverSampling. Finally, we applied tokenization and padding to the text sequences.

4.2 Feature Extraction

To capture meaningful features we used Term Frequency-Inverse Document Frequency (TF-IDF) for ML models And for DL models, we performed two types of embeddings: CBOW Word2Vec embedding and pre-trained FastText Tamil embedding. These embeddings were used to transform input tokens into dense vector representations, capturing semantic word relationships. The pre-trained embeddings were fine-tuned during model training to enhance the model's understanding of the text.

4.3 Model Building

In our research, we explored a variety of ML and DL models.

4.3.1 ML models

We trained traditional ML models such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines and Multinomial Naive Bayes on TF-IDF features. The models identify patterns statistically but struggle to extract the complex contextual meanings of the sentiments.

4.3.2 DL models

The deep learning models implemented for this task include GRU (Sachin et al., 2020), BiLSTM (Xu et al., 2019), BiGRU (Xu et al., 2024), hybrid CNN-GRU (Adam and Setiawan, 2023) and CNN-BiLSTM (Liu et al., 2020). Each DL model was trained for 8 epochs with a batch size of 64. We also applied layer normalization, dropout and the Adam optimizer (Kingma and Ba, 2014) for better generalization and more balanced learning. These models learn contextual word meanings based on how words appear in the training data via the embedding layer. The embedded vector is then passed into the network layers which capture contextual dependencies by learning the order and relationships between words and sentiment patterns.

5 Results and Discussion

In this section, we compare the performance achieved by different ML and DL models. The effectiveness of the models is primarily assessed based on the macro F1-score. The hyperparameters of the DL models were manually fine-tuned based on their performance on the validation data. The final hyperparameter values are as shown in Table 2. A summary of the precision (P), recall (R), and macro-F1 (MF1) scores for each model on the test set is presented in Table 3. Through our analysis, we found that the Bidirectional Gated Recurrent Unit (BiGRU) achieves the highest macro-F1 score of 0.33 on the test dataset using the FastText embedding, outperforming other ML and DL models. By processing both preceding and succeeding words, BiGRU enhances feature extraction for better sentiment classification than GRU. While CNN-GRU and CNN-BiLSTM benefit from convolutional feature extraction, the CNN component processes text with fixed-size receptive fields, which may restrict the recurrent layers' ability to

Table 2: The hyperparameters in BiGRU model

Hyperparameters	Values
Embedding Dimension	300
Units	300
Dropout Rate	0.2
Learning Rate	0.001
Optimizer	Adam
Loss Function	Sparse CCE
Batch Size	64
Epochs	8

Table 3: Results of various models on the test dataset

Classifier	P	R	MF1
LR	0.17	0.22	0.16
DT	0.21	0.23	0.21
RF	0.30	0.26	0.26
SVM	0.18	0.23	0.17
MNB	0.16	0.24	0.18
CBOW Embedding			
GRU	0.36	0.29	0.30
BiLSTM	**0.37**	0.29	0.30
CNN-BiLSTM	0.31	0.26	0.28
CNN-GRU	0.28	0.27	0.26
BiGRU	0.35	0.29	0.30
FastText Embedding			
GRU	0.32	0.29	0.30
BiLSTM	0.31	0.27	0.29
CNN-BiLSTM	0.34	0.31	0.31
CNN-GRU	0.27	0.27	0.26
BiGRU	0.35	**0.32**	**0.33**

capture long-range dependencies effectively. Moreover, these hybrid models have a higher number of parameters, making them more prone to overfitting. BiLSTM, though similar to BiGRU, has higher computational complexity and may overfit on smaller datasets, whereas BiGRU achieves a balance between performance and efficiency.

5.1 Quantitative Discussion

The performance of the BiGRU model for Political Multiclass Sentiment Analysis of Tamil X Comments is evaluated using a confusion matrix and ROC curve, as illustrated in Figure 3 and 4. The confusion matrix shows the model correctly classifies Opinionated comments with high accuracy (104 instances). However, Negative (16 misclassified as Opinionated), Neutral (25 misclassified as Opinionated), and Sarcastic (39 misclassified as Opinionated) sentiments exhibit significant misclassification, suggesting the model struggles to differentiate these classes. None of the above classes achieves a high correct classification rate (18 instances out of 23). The ROC curve highlights the model's varying discrimination ability. None of the above has the highest AUC (0.980), indicating strong separability, while Negative has the lowest (0.479), suggesting frequent misclassification. Other classes fall within 0.552–0.665, reflecting moderate distinction. The micro-average AUC of 0.700 suggests overall moderate performance, with challenges in handling nuanced sentiments.

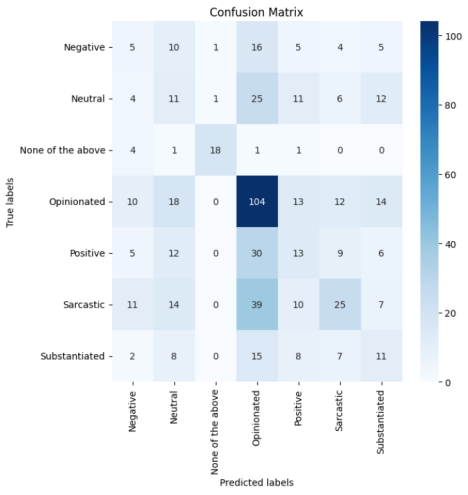


Figure 3: Confusion matrix of BiGRU model using FastText embedding

5.2 Qualitative Discussion

The BiGRU model's performance highlights the challenges posed by dataset imbalance, with minority classes like Negative, Substantiated, and Sarcastic often misclassified as Opinionated, reflecting a bias toward dominant sentiment patterns. Data augmentation played a crucial role in improving the model's performance by introducing more representative samples for the minority classes. Synonym replacement enhanced lexical diversity, while back-translation helped the model better capture variations in Negative, Substantiated, and None of the above sentiments. This led to a more balanced learning process, reducing extreme misclassification. However, the model still struggled with subtle sentiment distinctions, particularly implicit negativity and sarcasm, due to its reliance on sequential dependencies, which limited its ability to fully capture complex political context.

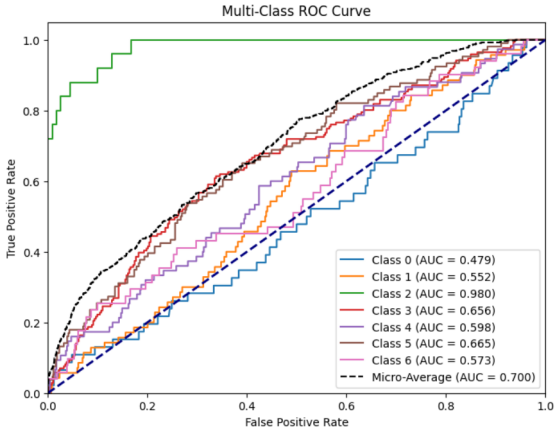


Figure 4: ROC curve of BiGRU model using FastText embedding

6 Conclusion

This study explored Political Multiclass Sentiment Analysis of Tamil X Comments as part of the DravidianLangTech@NAACL 2025 shared task. Key challenges included dataset imbalance and a test set that was not a strong representative of the embeddings. To address these issues, we employed synonym replacement to expand the dataset, improving the representation of embeddings and back-translation augmentation for underrepresented classes to enhance model robustness. Among various ML and DL models, the BiGRU model demonstrated the best performance, achieving an MF1 score of 0.33, a precision of 0.35, and recall of 0.32 using FastText embedding. Future work should explore domain-adaptive transformer models tailored for low-resource languages to further improve sentiment classification performance. Models such as mBERT, IndicBERT, and TamilBERT could be fine-tuned to political discourse data to enhance sentiment classification accuracy.

Limitations

The BiGRU model exhibited several limitations due to its reliance on sequential dependencies, which limited its ability to capture complex contextual nuances, leading to frequent misclassification of minority classes despite data augmentation efforts. Synonym replacement was applied to bring diversity to the training data, but it could not fully address the intricacies of sentiment variations. Back translation was implemented to improve prediction for minority classes, but this technique also struggled with the challenge of handling Out-of-Vocabulary (OOV) words, especially in informal social media text with transliterations, code-mixing, and spelling variations. Another key limitation was the imbalance between the training and test datasets, with the test dataset not being a strong representative of the training data, affecting the model's generalization ability. A more balanced dataset and transformer-based models could enhance contextual understanding and improve accuracy, particularly in handling nuanced sentiments and linguistic variations.

References

Ahmad Zahri Ruhban Adam and Erwin Budi Setiawan. 2023. Social media sentiment analysis using convolutional neural network (cnn) dan gated recurrent unit (gru). *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, 9(1):119–131.

Prathvi B, Manavi K, Subrahmanyapoojary K, Asha Hegde, Kavya G, and Hosahalli Shashirekha. 2024. MUCS@DravidianLangTech-2024: A grid search approach to explore sentiment analysis in code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 257–261, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Usha Devi Gandhi, Priyan Malarvizhi Kumar, Gokulnath Chandra Babu, and Gayathri Karthick. 2021. Sentiment analysis on twitter data by using convolutional neural network (cnn) and long short term memory (lstm). *Wireless Personal Communications*, pages 1–10.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, SK Lavanya, Durairaj Thenmozhi, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71.

T. Ananth Kumar Ahmad Jalili Mehdi Gheisari Yasir Malik Hsing-Chung Chen K. Suresh Kumar, A.S. Radha Mani and Ata Jahangir Moshayedi. 2024. Sentiment analysis of short texts using svms and vsms-based multiclass semantic classification. *Applied Artificial Intelligence*, 38(1):2321555.

Selam Kanta and Grigori Sidorov. 2023. Selam@ dravidianlangtech: Sentiment analysis of code-mixed dravidian texts using svm classification. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 176–179.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lakshmi Revathi Krosuri and Rama Satish Aravapalli. 2024. Novel heuristic bidirectional-recurrent neural network framework for multiclass sentiment analysis classification using coot optimization. *Multimedia Tools and Applications*, 83(5):13637–13657.

Zi-xian Liu, De-gan Zhang, Gu-zhao Luo, Ming Lian, and Bing Liu. 2020. A new method of emotional analysis based on cnn–bilstm hybrid neural network. *Cluster Computing*, 23:2901–2913.

Asrita Venkata Mandalam and Yashvardhan Sharma. 2021. Sentiment analysis of dravidian code mixed data. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 46–54.

Muhammad Kashif Nazir, CM Nadeem Faisal, Muhammad Asif Habib, and Haseeb Ahmad. 2025. Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages. *IEEE Access*.

K Nithya, S Sathyapriya, M Sulochana, S Thaarini, and CR Dhivyaa. 2022. Deep learning based analysis on code-mixed tamil text for sentiment classification with pre-trained ulmfit. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1112–1116. IEEE.

Kishore Kumar Ponnusamy, Charmathi Rajkumar, Prasanna Kumar Kumaresan, Elizabeth Sherly, and Ruba Priyadharshini. 2023. Vel@ dravidianlangtech: Sentiment analysis of tamil and tulu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 211–216.

M Rajasekar and Angelina Geetha. 2023. Sentiment analysis of tamil tweets using deep convolution neural networks. In *2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI)*, pages 1–5. IEEE.

Suba Sri Ramesh Babu. 2022. *Sentiment Analysis In Tamil Language Using Hybrid Deep Learning Approach*. Ph.D. thesis, Dublin, National College of Ireland.

Pradeep Kumar Roy and Abhinav Kumar. 2021. Sentiment analysis on tamil code-mixed text using bi-lstm. In *FIRE (Working Notes)*, pages 1044–1050.

Sharat Sachin, Abha Tripathi, Navya Mahajan, Shivani Aggarwal, and Preeti Nagrath. 2020. Sentiment analysis using gated recurrent neural networks. *SN Computer Science*, 1:1–13.

Lavanya Sambath Kumar, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 2024. Overview of second shared task on sentiment analysis in code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 62–70, St. Julian's, Malta. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022a. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech & Language*, 76:101407.

Kogilavani Shanmugavadivel, VE Sathishkumar, Sandhiya Raja, T Bheema Lingaiah, S Neelakandan, and Malliga Subramanian. 2022b. Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. *Scientific Reports*, 12(1):21557.

Poorvi Shetty. 2023. Poorvi@ dravidianlangtech: Sentiment analysis on code-mixed tulu and tamil corpus. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 124–132.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53. IEEE.

Farjana Alam Tofa, Lorin Tasnim Zeba, Md Osama, and Ashim Dey. 2025. CUET_INSights@NLU of Devanagari script languages 2025: Leveraging transformer-based models for target identification in hate speech. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 267–272, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Guixian Xu, Yueting Meng, Xiaoyu Qiu, Ziheng Yu, and Xu Wu. 2019. Sentiment analysis of comment texts based on bilstm. *Ieee Access*, 7:51522–51532.

Wei Xu, Jianlong Chen, Zhicheng Ding, and Jinyin Wang. 2024. Text sentiment analysis and classification based on bidirectional gated recurrent units (grus) model. *arXiv preprint arXiv:2404.17123*.

CIC-NLP@DravidianLangTech 2025: Detecting AI-generated Product Reviews in Dravidian Languages

Tewodros Achamaleh[1], Abiola T. O.[1], Lemlem Eyob[1], Mikiyas Mebiratu[2], Grigori Sidorov[1]
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico
[2]Wolkite University, Department of Information Technology, Wolkite, Ethiopia

Abstract

AI-generated text now matches human writing so well that telling them apart is very difficult. Our CIC-NLP team submits results for the DravidianLangTech@NAACL 2025 shared task to reveal AI-generated product reviews in Dravidian languages. We performed a binary classification task with XLM-RoBERTa-Base using the DravidianLangTech@NAACL 2025 datasets offered by the event organizers. By training the model effectively, our experiments distinguished between human and AI-generated reviews with scores of 0.96 for Tamil and 0.88 for Malayalam in the evaluation test set. This paper presents detailed information about preprocessing, model architecture, hyperparameter fine-tuning settings, the experimental process, and the results. The source code is available on GitHub.[1]

1 Introduction

The fast growth of Large Language Models (LLMs) now changes how natural language processing works across many uses (Yigezu and Tesfaye, 2023; Kolesnikova and Ivanov, 2023; Adebanji and Okoro, 2024; García-Vázquez and Rodriguez, 2023; Laureano and Calvo, 2024; Aguilar-Canto and Ramirez, 2023; Ojo and Bello, 2024; Brown and Leike, 2023; Abiola et al., 2025b,a). Computer algorithms make Machine-generated text through AI while showing human writing ability with limited human involvement. MGT has revolutionized production through automated content creation, but labs now must excel at recognizing MGT from HWT texts, especially in situations demanding proof like product evaluation.

Human authors create text that harnesses personal experiences to comprehend cultures and emotions, which allows them to present detailed feelings that fit perfectly into their context. According to (Zhang et al., 2024), MGT shows language precision at its surface level but fails to achieve the contextual synergy present in HWT. Underrepresented Dravidian languages show distinct characteristics that make their interpretation different from other languages (Conneau et al., 2020; Ruder et al., 2023).

Detecting MGT is essential for stopping online lies and resolving ethical issues with AI-generated content (Ansarullah, 2024; Floridi and Cowls, 2023). Language models built at scale need training data that holds stereotypes to produce outputs that follow established verbalization patterns(Gallegos et al., 2024; Brennan and Greenstadt, 2023). These computing system prejudices create analytical opportunities to tell actual human-written text from machine-generated text through detailed language marker inspection. Our team joins the DravidianLangTech@NAACL 2025 Shared Task to create AI-generated product review detection systems for Tamil and Malayalam. Our work involved differentiating AI-generated and human-written reviews across Tamil and Malayalam using an exceptional data resource that includes multiple language forms from human writers and computer systems. Our research used XLM-RoBERTa-Base, a transformer model for multilingual text understanding (Liu and Ott, 2023), as the basis for our experiment. Our research confirms how the model understands varied language styles and shows why different data sets need separate treatment in AI content detection technology.

Our methodology achieved macro average F1 scores of 00.96 for Tamil and 0.88 for Malayalam on the evaluation test set. This paper's main contribution is to provide insights into preprocessing, model architecture, hyperparameter tuning, and evaluation. It correspondingly contributes to the growing AI content detection research in low-resource languages. Our work brings to the forefront the potential of fine-tuned multilingual

[1]https://github.com/teddymas95/AI-generated-Product-Reviews

models for NLP in underrepresented languages by addressing the complexity of multilingual AI-generated texts.

2 Related Work

From here to the research, the focus was on finding clear indicators of AI-generated content through pattern detection or verbalization inconsistencies (Maimone and Jolley, 2023; Aydin and Kara, 2023; Clark et al., 2023). Nevertheless, with the development of generative models regarding text generation quality and contextual coherence (Smith et al., 2023; Brown et al., 2024), machine-generated text increasingly became more complicated to differentiate. As these advancements were made, traditional rule-based systems became not adequate, pushing us into the field of deep learning approaches, in particular using transformer-based models (Kierner et al., 2023; Chen and Wang, 2024; Jurafsky and Martin, 2023). Natural language processing (NLP) has come a long way, but transformer models have greatly improved it. Several studies have shown them to be very strong at NLP tasks such as sentiment analysis, text classification, and data summarization (Soto et al., 2024; Hoang, 2024; Zhang et al., 2024; Ruder et al., 2023).

(Gupta and Verma, 2023) XLM-RoBERTa was consistently widely preferred for Multilingual tasks, especially in low-resource languages like Tamil and Malayalam, due to its strong cross-lingual performance (Conneau et al., 2020). In particular, (Li et al., 2024) studied the issue of how to build robust AI detection systems for varied text types and multiple language models. The paper emphasized the need to deal with text variability in real-world text and showed how named entities and structural details may help identify differences between AI-generated and human-written text, but with slight differences as AI systems improve (Brennan et al., 2023). (Fernández-Hernández et al., 2023; Eyob et al., 2024) Performed a series of experiments with multilingual BERT for the AuTexTification shared task at IberLEF 2023 concerning distinguishing AI-generated texts (García-Vázquez and Rodriguez, 2023). Their findings demonstrated that fine-tuned transformer models could outperform traditional machine-learning techniques without including metadata features like readability and sentiment.

Also, (Kumar et al., 2024) looked at how well hybrid transformer-based architectures deal with linguistic diversity, specifically in classifying texts from several domains (Joshi et al., 2024). These studies have been restricted to high-resource languages, and it remains challenging to detect machine-generated text (MGT) in low-resource languages like Tamil and Malayalam (Chatterjee et al., 2023; Kumar et al., 2023; Achamaleh et al., 2024). As linguistically diverse and morphologically complex as Hindi is, and scarce are large annotated datasets, tailored methods are called. This work proceeds prior work using XLM-Roberta in its multilingual capabilities, developing the detection of AI-generated content in Dravidian languages and outperforming state of the art.

3 AI vs. Human Text Detection

3.1 Dataset Analysis

The organizers provided datasets for training and testing data through Google Drive(Premjith et al., 2025). Each dataset consists of the following columns: ID, DATA, and LABEL. The Label column contains two values: The datasets classify text as HUMAN when humans compose it and AI when AI systems produce it. Our primary objective is to differentiate AI-generated text from human-written content. The Tamil dataset includes 808 records of AI-generated (405 texts) and human-written material (403 texts). The Malayalam dataset provides 800 texts made by both AI generators and humans, with 400 texts in each group. The team made this dataset to represent normal content variations in real-world data, supporting high-quality model testing and training. During this task, the datasets were split into training, validation, and testing sets, enabling the fine-tuning of our XLM-RoBERTa-Base model. The balanced class distribution in the datasets contributed to achieving reliable and unbiased model performance across Tamil and Malayalam.

3.2 XLM-RoBERTa-Base

We used the Transformers Library from Hugging Face and fine-tuned the XLM-RoBERTa Base, a multilingual transformer model for binary Tamil and Malayalam text classification. To prepare and format the dataset and satisfy the model's input needs, we respected specific tokenization and inherent linguistic caveats about these languages. Our team then processed the dataset by passing it through the XLM-RoBERTa tokenizer to prepare for training and testing. Our method included

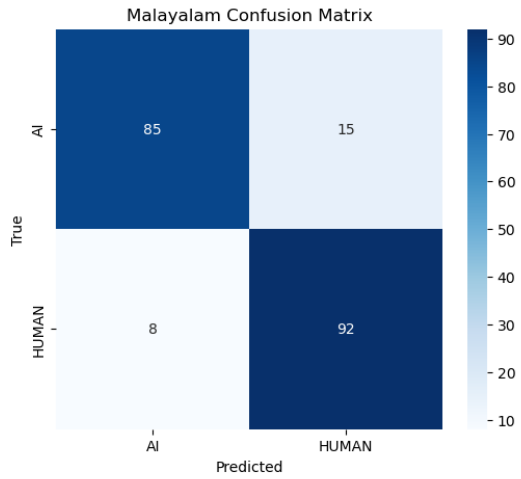


Figure 1: Malayalam Confusion Matrix

adding dropout to prevent model overtraining and saving checkpoints during training to pick the version that did best according to validation metrics. We adopted a cosine-annealing learning rate schedule to stabilize training and improve the final model performance. This paper provides the methodology of adapting to use XLM-RoBERTa-Base effectively for this task. Section 4 explains how we preprocessed the dataset, encoded the text, and were ready for further analysis.

4 System Setup and Experiments

4.1 System Setup

We trained the XLM-RoBERTa-Base model as a multilingual transformer architecture tuned to identify pairs of binary classes in Tamil and Malayalam language datasets. The datasets were preliminary processed by Hugging Face AutoTokenizer, which turned text entries into model-ready tokenized content. Hugging Face datasets library divided our information into 90% training data and 10% test data sets. To recognize text types, the XLM-RoBERTa model required the addition of a classification segment that generated human or AI predictions. With a learning rate of 3e-5, we trained our model across five epochs using batches of 8 and regularized dropout layers to avoid overfitting. Our system selected the optimal model results by evaluating performance on early stop conditions and checkpointed models. We assessed model performance using F1 scores, precision, recall, and accuracy during test dataset predictions.

4.2 Experiments

For the binary classification of human- and AI-generated text in Tamil and Malayalam datasets, we fine-tuned the XLM-RoBERTa-Base model. This multilingual transformer design took in both dataset characteristics well. The model achieved better results by selecting specific values for important training settings such as batch size, learning rate, and training steps. Our model used GPU computing during five training epochs and divided gradient updates into two steps to fit memory. We took advantage of the mixed precision training to make training run faster and to prevent overfitting by early stopping based on the validation of the F1 score. Moreover, a cosine-annealing learning rate scheduler and warmup steps stabilized the training, with the learning rate starting low and increasing gradually at the beginning and becoming lower after some time.

DataCollatorWithPadding was used to dynamically pad input sequences for each batch to the maximum sequence length for computational efficiency. This reduced the number of extra operations on padding tokens, which made the model more attentive to meaningful text content. F1-score and loss metrics were used closely to indicate the training and validation performance. The results include training and validation plots, which show that the model achieved competitive performance with macro F1 scores. The results indicate the robustness of the fine-tuning approach and the selection of good hyperparameters. After each epoch, we evaluated the model's performance on the development dataset, tracking its progress and ensuring the training and validation metrics were aligned. This helped identify potential issues such as overfitting or underfitting early in the process.

5 Results

Our evaluation tests the performance of our fine-tuned XLM-RoBERTa-Base model across Tamil and Malayalam datasets for a binary classification setup. We evaluated model performance by running text predictions on development data and measured accuracy plus micro and macro F1 scores. Our Tamil model achieved 0.96 accuracy as measured by macro F1 scores to differentiate content created by AI from human producers. The dataset balance and rich vocabulary influenced Tamil text, giving rise to this excellent model performance. Even with uneven class distribution in Malayalam data,

Language	Model	Precision	Recall	F1-Score	Accuracy
Malayalam	xlm-roberta-base	0.9739	0.9739	**0.9739**	0.975
	distilbert-base-uncased	0.9194	0.9271	0.9226	0.925
	bert-base-multilingual-cased	0.9479	0.9479	0.9479	0.950
Tamil	xlm-roberta-base	0.9509	0.9286	**0.9358**	0.9383
	distilbert-base-uncased	0.9423	0.9143	0.9225	0.9259
	bert-base-multilingual-cased	0.9509	0.9286	0.9258	0.9283

Table 1: Model Comparison for Malayalam and Tamil on the Development Dataset.

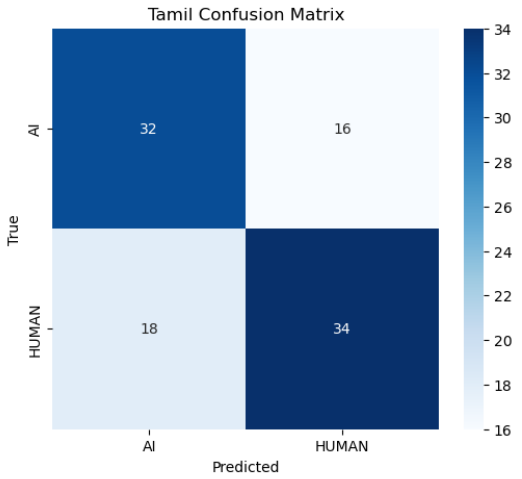


Figure 2: Tamil Confusion Matrix

the model delivered an impressive 0.88 Macro F1 score, demonstrating its ability to work across multiple languages.The model shows a strong ability to differentiate AI and human-generated text in Tamil and Malayalam with reliable detection accuracy.

6 Discussion

The detection of AI-generated reviews using XLM-RoBERTa in Tamil and Malayalam was highly effective. Training data balance for the Tamil model strengthened its generalization capability. The class imbalance in Malayalam did not affect its ability to maintain high generalization performance. Achieving multilingual NLP success depends on synchronous data quality management and proper class balancing capabilities, making transformer models ideal for low-resource language processing.

XLM-RoBERTa outperformed DistilBERT and BERT-multilingual in precision and F1-score. DistilBERT was efficient but misclassified many authentic reviews, while BERT-multilingual had uneven results, especially with Malayalam. XLM-RoBERTa showed reliable performance, though all models struggled with unclear cases. Future improvements could include domain-specific training and features like readability scores and syntactic analysis. Table 1 compares the models for both languages.

6.1 Error Analysis

The minority classes in the imbalanced Malayalam dataset showed most of the misinterpreted classification labels. AI-generated reviews in Tamil easily fooled human scrutiny because they were presented as if written by human writers. The number of wrong classifications in Malayalam increased because the language uses intricate sentence formats and blends two different written systems. Research results indicate that it is necessary to improve algorithmic models by introducing language elements that exceed simple token recognition processes. Figures 1 and 2 show the confusion matrix.

Conclusion

The research examined the power of transformer-based models to find AI-generated product reviews across the two Dravidian languages, Tamil and Malayalam. The XLM-RoBERTa model achieved better results, particularly in Tamil, since its balanced dataset helped it improve generalization abilities. The Malayalam model demonstrated robustness even though its performance was affected by the class imbalance problem. The analysis of misclassification errors during testing showed that AI mistaken instances mainly occurred when minority classes contained text similar to actual human writing. XLM-RoBERTa performed best among all three models during comparison tests because it delivered maximum precision and F1-score measurements for both language codes. All produced models encountered difficulties when classifying ambiguous instances, suggesting enhanced improvements through linguistic features must be implemented. Properly selecting high-quality multilin-

gual datasets plays a critical role in successful NLP tasks. The future development of AI-generated text detection in low-resource languages requires research on syntactic feature integration, semantic feature integration, domain-specific fine-tuning, and metadata-based improvement methods.

Limitations

This study faced several challenges because class imbalance negatively influenced the performance of the Malayalam model. The model encountered difficulties with generalization because the dataset had a limited capacity to handle different writing styles. The models failed to function correctly while processing ambiguous cases with AI text similar to human writing. Future improvements must concentrate on growing more enormous datasets with balanced distribution and developing advanced linguistic elements to improve detection precision levels.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Tolulope O. Abiola, Tewodros A. Bizuneh, Oluwatobi J. Abiola, Temitope O. Oladepo, Olumide E. Ojo, Adebanji O. O., Grigori Sidorov, and Olga Kolesnikova. 2025a. Cic-nlp at genai detection task 1: Leveraging distilbert for detecting machine-generated text in english. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Tolulope O. Abiola, Tewodros A. Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide E. Ojo. 2025b. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Tewodros Achamaleh, Lemlem Kawo, Ildar Batyrshini, and Grigori Sidorov. 2024. Tewodros@ dravidian-langtech 2024: Hate speech recognition in telugu codemixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 96–100.

Tunde Adebanji and Chima Okoro. 2024. Ethical implications of ai in generating human-like text. *AI Society*, 39:567–580.

Diego Aguilar-Canto and Sofia Ramirez. 2023. Challenges in detecting machine-generated text in under-resourced languages. *Language Resources and Evaluation*, 57:145–161.

M. et al. Ansarullah. 2024. Inceptor regulates insulin homeostasis. *Nature Metabolism*. Referencing Gallegos et al., 2024.

Mehmet Aydin and Elif Kara. 2023. Advancements in detecting ai-generated texts: Challenges and methodologies. *AI and Society*.

M. Brennan, S. Afroz, and R. Greenstadt. 2023. Forensic linguistics for ai text detection.

M. Brennan and R. Greenstadt. 2023. Linguistic markers for ai text detection.

T. Brown and J. Leike. 2023. The sociotechnical impact of large language models.

T. Brown, B. Mann, and N. Ryder. 2024. Gpt-4: Scaling generative text quality.

Rahul Chatterjee et al. 2023. Challenges in nlp for low-resource languages: A focus on tamil and malayalam. *Low-Resource NLP Journal*.

Yuxin Chen and Jie Wang. 2024. Transformer-based architectures in modern nlp. *NLP Research Journal*.

E. Clark, A. Gupta, and K. Lee. 2023. Early detection methods for ai-generated text.

A. Conneau, K. Khandelwal, and N. Goyal. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Lemlem Eyob, Tewodros Achamaleh, Muhammad Tayyab, Grigori Sidorov, and Ildar Batyrshin. 2024. Stress recognition in code-mixed social media texts using machine learning. *International Journal of Combinatorial Optimization Problems and Informatics*, 15(1):32.

Ana Fernández-Hernández et al. 2023. Autextification shared task at iberlef 2023: Experiments with multilingual bert for ai text detection. In *Proceedings of IberLEF 2023*. Springer.

L. Floridi and J. Cowls. 2023. Ethical governance of generative ai.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Elena García-Vázquez and Pedro Rodriguez. 2023. Bias in machine-generated text: A case study on multilingual models. *International Journal of Artificial Intelligence*, 32:45–62.

Rishi Gupta and Ananya Verma. 2023. Cross-lingual applications of xlm-roberta in low-resource nlp tasks. *Journal of Computational Linguistics*.

T. Hoang. 2024. Transformer models in multilingual nlp. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 456–465.

P. Joshi, S. Santy, and A. Budhiraja. 2024. Challenges in low-resource language nlp.

D. Jurafsky and J. Martin. 2023. From rule-based systems to transformers: A survey of nlp paradigms.

Tobias Kierner et al. 2023. Transformer models in natural language processing: A survey of advancements and applications. *Computational Linguistics Today*.

Anna Kolesnikova and Sergey Ivanov. 2023. Exploring multilingual text representations with transformer models. *Transactions of the ACL*, 11:212–230.

Arjun Kumar et al. 2024. Hybrid transformer-based architectures for multilingual text classification. *Journal of Artificial Intelligence and Language Technologies*.

R. Kumar, S. Murugesan, and B. Rajendran. 2023. Dravidian language processing: Trends and gaps.

Miguel Laureano and Ana Calvo. 2024. Language models and their role in sentiment analysis. *Journal of Sentiment Analysis*, 18:321–337.

Wei Li et al. 2024. Challenges in ai-detection systems for multilingual text classification. *Multilingual AI Journal*.

Y. Liu and M. Ott. 2023. Xlm-roberta for multilingual understanding.

John Maimone and Sarah Jolley. 2023. Identifying ai-generated content: Pattern detection and verbalization inconsistencies. *Journal of Artificial Intelligence Research*.

Adewale Ojo and Fatima Bello. 2024. Cross-lingual learning: Advancements in text classification. *Journal of Cross-lingual NLP*, 25:89–105.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

S. Ruder, M. Peters, and S. Swayamdipta. 2023. Domain adaptation in transformer models.

Hannah Smith et al. 2023. Overcoming the challenges of detecting advanced ai-generated text. *Journal of Machine Learning Applications*.

R. Soto et al. 2024. Applications of xlm-roberta in multilingual text analysis. *Transactions on Computational Linguistics*, 12:234–245.

Alemayehu Yigezu and Meron Tesfaye. 2023. The advancements in cross-lingual nlp applications. *Journal of Computational Linguistics*, 49:102–119.

Wei Zhang et al. 2024. Evaluation of transformer models in multilingual sentiment analysis tasks. *Sentiment Analytics Quarterly*.

One_by_zero@DravidianLangTech 2025: A Multimodal Approach for Misogyny Meme Detection in Malayalam Leveraging Visual and Textual Features

Dola Chakraborty*, **Shamima Afroz Mithi***
Jawad Hossain and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904012, u1904106, u1704039}@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

Misogyny memes are a form of online content that spreads harmful and damaging ideas about women. By combining images and text, they often aim to mock, disrespect, or insult women, sometimes overtly and other times in more subtle, insidious ways. Detecting Misogyny memes is crucial for fostering safer and more respectful online communities. While extensive research has been conducted on high-resource languages (HRLs) like English, low-resource languages (LRLs) such as Dravidian (e.g. Malayalam) remain largely overlooked. The shared task on Misogyny Meme Detection, organized as part of DravidianLangTech@NAACL 2025, provided a platform to tackle the challenge of identifying misogynistic content in memes, specifically in Malayalam. We participated in the competition and adopted a multimodal approach to contribute to this effort. For image analysis, we employed a ResNet18 model to extract visual features, while for text analysis, we utilized the IndicBERT model. Our system achieved an impressive F1-score of 0.87, earning us the 3rd rank in the task.

1 Introduction

Misogynistic memes have a significant influence as they contribute to normalizing and perpetuating harmful attitudes and behaviors (Paciello et al., 2021). These memes use social media's visual-textual and viral qualities to quietly embed and disseminate sexist beliefs, frequently making identification and intervention challenging. Detecting misogyny in memes poses unique challenges due to their multimodal nature. Sometimes text and image individually exhibit no offense, but combining both elements can convey implicit or contextual misogyny, making the meme offensive when taken as a whole (Chen et al., 2024; Gasparini et al., 2022). While substantial research has been conducted on misogyny detection in high-resource languages like English (Fersini et al., 2022), low-resource languages, particularly Dravidian languages such as Malayalam, remain underexplored. Malayalam, a Dravidian language spoken predominantly in the Indian state of Kerala, faces a growing issue of misogynistic memes on digital platforms, highlighting the need for targeted research.

Misogynistic memes in Malayalam often present additional challenges due to linguistic nuances, transliterated text (Malayalam written in English script), and the interplay of regional cultural references. The issue is made worse by the dearth of extensively annotated datasets in Malayalam, which makes it more difficult to create reliable detection methods. As participants in this shared task, our work makes the following notable contributions:

- Proposed a transformer-based approach to classify Malayalam misogynistic memes as *Misogynistic (Miso)* or *Non-misogynistic (NMiso)*.
- Experimented with several DL, transformer-based models to extract visual and textual features and employed late fusion to combine features from both modalities to detect misogynistic memes.

2 Related Work

Shushkevich and Cardiff (2018) analyzed tweets from Twitter for the Automatic Misogyny Identification (AMI) task at EVALITA 2018 and achieved an F1 score of 0.78 using Logistic Regression (LR) Devi and Saharia (2021) focused on misogyny detection in English, classifying texts as misogynous or not. They achieved 93.43(%) accuracy using a Bi-LSTM model. Goenaga et al. (2018) was part of the AMI-IberEval 2018 competition, identifying misogyny in English and Spanish tweets. They used a Bi-LSTM with CRF, achiev-

*Authors contributed equally to this work.

ing 78.9(%) accuracy for English and 76.8(%) accuracy for Spanish. Srivastava (2022) participated in SemEval-2022 Task 5, specifically SubTask-A, which focused on identifying whether a meme contained misogyny. Using the ResNet-50nsfw model, they achieved a notable 7th-place ranking with an F1 score of 0.759. Rizzi et al. (2023) focused on misogyny detection in memes using unimodal and multimodal approaches, achieving an overall accuracy of 61.43(%). Their method incorporated a bias mitigation strategy based on Bayesian Optimization to improve model performance. Chinivar et al. (2024) tackled misogynistic meme detection using multimodal models on a benchmarked dataset. Their approach combined XLM-R for text and Swin for images, resulting in an F1 score of 0.7607. Singh et al. (2024) performed binary and multi-label classification of misogynistic memes. Their top-performing model, BiT (image) + MuRIL (text), for binary classification, achieved a high F1 score of 0.7319.The task described in Arango et al. (2022) involved identifying misogynous memes for the Multimedia Automatic Misogyny Identification (MAMI) task at SemEval-2022. Using a multimodal system based on the CLIP model, they achieved an F1 score of 71(%). Raha et al. (2022) addressed misogyny detection in memes as a binary classification task. Their best-performing models, VisualBERT and ViLBERT, attained an F1 score of 0.712. Using the MAMI task dataset, Ravagli and Vaiani (2022) worked on identifying misogynistic memes for SemEval-2022 Task 5 (MAMI). They combined Mask R-CNN for image processing and VisualBERT for multimodal processing. The VisualBERT (COCO) model achieved an F1 score of 0.670. Chen and Pan (2022) conducted hateful meme detection through text and image analysis. Their approach used OSCAR+RF, integrating the OSCAR Vision-Language Pre-Training Model with a Random Forest classifier, achieving an accuracy of 0.684.

3 Task and Dataset Description

In this shared task, the focus is on misogyny meme detection in Malayalam language. The task involves classifying memes as Misogynistic (labeled as 1) or Non-misogynistic (labeled as 0) in the Malayalam language. The dataset (Ponnusamy et al., 2024; Chakravarthi et al., 2024, 2025) provided by the organizers is sourced from various social media platforms.

Table1 depicts the statistical distribution of data. The training dataset contains a total of 640 samples, with 259 labeled as misogynistic and 381 as non-misogynistic. The validation dataset consists of 160 samples, including 63 misogynistic and 97 non-misogynistic samples. Out of the 200 samples in the test dataset, 78 are classified as misogynistic and 122 as non-misogynistic.

Classes	Train	Valid	Test	T_W
Mis	259	63	78	6398
NMiso	381	97	122	11004
Total	640	160	200	17402

Table 1: Dataset Statistics for Train, Validation, and Test Sets. (T_W denotes total words)

The dataset is provided in the form of an image with an associated transcription. We utilized image, text and multimodal (text + image) features to address this task. The implementation of our proposed approach has been made publicly available, and the source code can be accessed on GitHub[1].

4 Methodology

Before adopting a multimodal approach, we have focused on exploiting the visual aspects of memes by developing several CNN architectures and a Vision Transformer. For the textual aspects, we have implemented Text-CNN, Malayalam BERT, and IndicBERT. Finally, the visual and textual features are combined using fusion techniques to enhance the model's performance in detecting misogynistic content.Figure 1 depicts a schematic process in detecting fake news, illustrating each major phase.

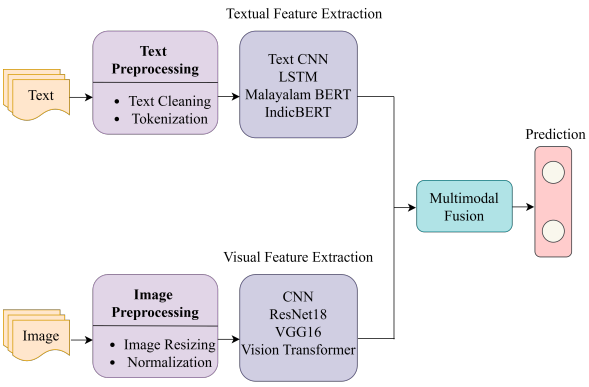


Figure 1: Schematic process of misogyny meme detection.

[1] https://github.com/DolaChakraborty12/Misogyny_Meme_Detection

4.1 Data Pre-processing

In the preprocessing step, unwanted symbols and punctuation were removed from the text. The text was then tokenized using a pre-trained BERT tokenizer, which converted the words into unique numerical representations. The sequences were padded to a fixed length of 128 tokens to ensure consistent input size across all samples. The images were resized to a fixed size of 224x224 pixels. In RGB format, the images were transformed to a size of (224x224x3). Each image was then normalized by scaling the pixel values to a range between 0 and 1. Additionally, image transformations were applied, including resizing, normalization, and conversion to tensors to prepare the data for input into the CNN model.

4.2 Visual Approach

For visual elements, several CNN-based architectures and transformer-based models were evaluated, such as Vision Transformer (ViT), VGG16, and a Convolutional Neural Network (CNN). The ViT was fine-tuned with a learning rate of 1e-5, batch size of 16, and 15 epochs, using AdamW optimizer and Binary Cross-Entropy loss. The VGG16 model was modified with custom layers and trained with a learning rate of 1e-5, batch size of 16, and 100 epochs, using categorical cross-entropy loss. Lastly, the CNN model, trained with a learning rate of 5e-5, batch size of 16, and 100 epochs, using categorical cross-entropy loss.

4.3 Textual Approach

In the textual approach, we used BiLSTM, TextCNN, LSTM + CNN, and Malayalam BERT.

- **BiLSTM**: Input text was tokenized using TensorFlow Keras (maximum vocabulary size: 10,000, input length: 100) and passed through an embedding layer (128-dimensional). It was then processed by a bidirectional LSTM layer with dropout rates of 0.8 and 0.5, followed by a dense layer with ReLU activation and L2 regularization (0.01). The model was trained using the Adam optimizer, binary cross-entropy loss, and balanced class weights, with a batch size of 16 for 10 epochs.

- **Malayalam BERT**: The model was fine-tuned using the Adam optimizer and binary cross-entropy loss with a batch size of 16 for 10 epochs, using a learning rate of 2e-5 and a weight decay of 0.01 to prevent overfitting.

- **Text-CNN**: Input text was first passed through an embedding layer (100-dimensional), followed by a convolutional layer with 128 filters and a kernel size of 5. The output was then processed by a max pooling layer, followed by a fully connected layer with 128 units and ReLU activation, and finally passed through a dropout layer (0.5) before the output layer. The model was trained using the Adam optimizer and binary cross-entropy loss with a batch size of 32 for 100 epochs.

- **LSTM+CNN**: Input text was first processed by an embedding layer and then passed through a bidirectional LSTM layer with 128 units. The output was then fed into a convolutional layer, followed by a max pooling layer. Finally, the processed features were passed through fully connected layers before the output layer. The model was trained using the Adam optimizer and binary cross-entropy loss with a batch size of 32 for 15 epochs.

4.4 Multimodal Approach

For visual feature extraction, multiple pretrained models, including Vision Transformer (ViT), CLIP, and ResNet-18 were utilized due to their strong performance in capturing diverse image features, ViT excels in global context understanding, CLIP aligns visual and textual features, and ResNet-18 excels in robust, hierarchical feature extraction. ViT processed the images by resizing them to 224x224 pixels and generating representations using a sequence-based transformer approach. CLIP and ResNet-18 were employed with batch sizes of 16 and learning rates set to 1e-5 and 2e-5 respectively, to extract additional visual and contextual features. The resulting image embeddings were passed through fully connected layers to reduce dimensionality and align with textual features for multimodal fusion.

For textual feature extraction, the system implemented advanced language models like Malayalam BERT and IndicBERT. These models are pretrained on large corpora of Dravidian languages, enabling them to capture language-specific syntactic and semantic features essential for accurately understanding the nuanced textual content in Malayalam memes. Malayalam BERT, fine-tuned for Malayalam text, was used to handle transliterated and native Malayalam text effectively. IndicBERT, being a multilingual model, was also implemented

for handling Malayalam effectively. Text inputs were tokenized, where sequences were padded or truncated to a fixed length, and embeddings were extracted from the model's output. The extracted embeddings were then passed through fully connected layers to transform them into a compact feature vector.

The outputs of both the visual and textual models were concatenated at a multimodal fusion layer. This integration of visual and textual features ensured that both modalities contribute effectively to the final prediction. A fully connected classification layer with a sigmoid activation was added after the fusion layer to produce the final binary prediction.

Training was conducted end-to-end with the binary cross-entropy loss function and the Adam optimizer, with a learning rate of 2e-5, a maximum sequence length of 128, and a batch size of 16. Table 2 demonstrates the hyperparameters of the best performed model(ResNet and Malayalam BERT).

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	2e-5
Batch Size	16
Max Length	128
Epochs	5

Table 2: Hyperparameter setup

5 Results and Analysis

Table 3 illustrates the performance of the various deep learning (DL) and transformer-based models employed on the test dataset across different approaches. The multimodal approach outperformed both visual and textual approaches. In the visual approach, the Vision Transformer (ViT) outperformed deep learning models, achieving an F1-score of 0.79. Malayalam BERT outperformed the other models in the textual method. Finally, in the multimodal approach, the fusion of ResNet (for visual features) and Malayalam BERT (for textual features) provided the best result, achieving a macro F1-score of 0.86.

6 Error Analysis

A detailed error analysis of the best-performed model is executed using quantitative and qualitative approaches.

Approach	Classifier	P	R	F1
Visual	ViT	0.98	0.56	0.79
	VGG16	0.77	0.67	0.68
	CNN	0.58	0.51	0.54
Textual	BiLSTM	0.73	0.72	0.72
	Malayalam BERT	0.41	0.80	0.54
	CNN	0.59	0.85	0.71
	LSTM + CNN	0.62	0.69	0.71
Multimodal	ResNet18 + Malayalam BERT	0.82	0.82	0.82
	ResNet18 + IndicBERT	**0.87**	**0.88**	**0.87**
	ViT + Malayalam BERT	0.88	0.79	0.81
	ViT + IndicBERT	0.86	0.86	0.86
	CLIP + Malayalam BERT	0.88	0.83	0.85

Table 3: Performance of various DL and Transformer-based models on the test set. P (Precision), R (Recall), F1 (macro F1-score).

Quantitative Analysis: Figure 2 presents the confusion matrix of the best-performing multimodal model, ResNet18 + IndicBERT. A detailed error analysis of the fine-tuned multimodal model is performed based on the confusion matrix. It is evident from the confusion matrix that, out of 200 samples, 176 are correctly predicted. The model misclassifies 11 misogynistic samples as non-misogynistic and 13 non-misogynistic samples as misogynistic.

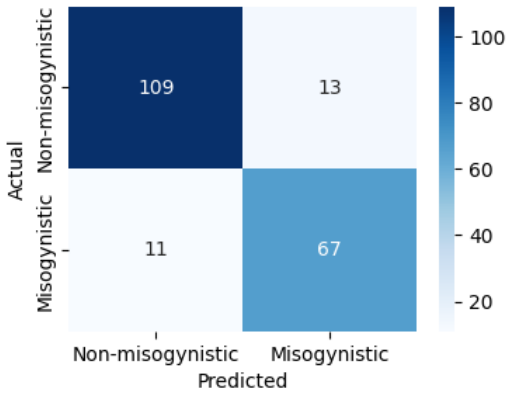


Figure 2: Confusion matrix of the best-performed model (ResNet18 + IndicBERT).

Qualitative Analysis: A comparison of actual labels and predicted labels for a particular transcription is illustrated in Figure 3. The first two samples are incorrectly predicted as misogynistic, even though they are non-misogynistic. However, the next two samples are predicted correctly as their actual labels.

The misclassifications observed in the results can be attributed to the challenges inherent in multimodal fusion, where both image and text features are integrated. While the fusion of deep learning-based image features (extracted via ResNet18) and

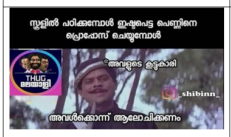
Images ID	Transcriptions	Predicted Labels	Actual Labels
	സ്കൂളിൽ പഠിക്കുമ്പോൾ ഇഷ്ടപെട്ട പെണ്ണിനെ പ്രപ്പോസ് ചെയ്യുമ്പോൾ "അവളുടെ കൂട്ടുകാരി അവൾക്കൊന്ന് ആലോചിക്കണം	1	0
	അറിയാൻ പാടില്ലാത്ത ഒരാളെകുറിച്ച് നമ്മൾ ഒരിക്കലും കുറ്റം പറയരുത് \nകാരണം അടുത്തറിയുമ്പോഴായിരിക്കും അവനൊരു പാവമാണെന്ന് പലർക്കും മനസ്സിലാവുക	1	0
	എന്തൊക്കെ ആകിയിട്ടും ഒരു മെന വരുന്നില്ലലോ നിഖിയേ..	1	1
	ഇത് ഞാൻ ചെറുതായിരുന്നപ്പോൾ ഇത് 5ആം ക്ലാസ്സ് വരെ കണ്ടു പിന്നെ ഇത്. ഇത്... ഇന്ന് രാവിലെയും കൂടെ കണ്ടു	0	0

Figure 3: Some predicted outputs by (ResNet18 + IndicBERT).

transformer-based text features (from IndicBERT) enables the model to leverage complementary information, it may also introduce certain ambiguities. The concatenation of these two distinct feature types can sometimes lead to confusion in classification, as the model must balance the influence of both modalities. The model often struggles with sarcastic statements where the textual content appears non-misogynistic but carries misogynistic undertones when paired with the image. The model fails to capture such implicit misogyny, leading to misclassification. Some memes contain misleading or ambiguous text, cultural references, slang, and region-specific humor, particularly in Malayalam, where proverbs or idiomatic expressions may have context-dependent misogynistic intent, challenging the model's classification.

7 Conclusion

This work explored the effectiveness of various DL and transformer-based models for misogyny meme detection in Malayalam. Different modalities, including textual, visual, and multimodal approaches, are systematically evaluated. ViT achieved the highest F1 score among the visual models, demonstrating its ability to capture complex visual patterns. BiLSTM outperformed other models for the textual modality, showcasing its strength in handling sequential data. However, the best overall performance is achieved through a multimodal approach that combined ResNet18 and IndicBERT, resulting in the highest F1 score of 0.87. This result highlights the significance of integrating complementary features from textual and visual modalities for addressing challenging tasks like misogyny meme detection. Future work can enhance this task by incorporating larger datasets to improve model robustness, reduce bias, and enhance generalization by exposing the model to a diverse range of misogynistic and non-misogynistic memes. Additionally, exploring single transformer-based approaches for multimodal learning and investigating large language models can improve performance. While this study focuses on Malayalam memes, our results suggest that IndicBERT outperforms language-specific models like MalayalamBERT, highlighting its potential for cross-lingual effectiveness in other Dravidian languages.

Limitations

The current model posses several weaknesses. A few of them are illustrating in the following:

- Combining ResNet18 and IndicBERT features can introduce confusion, leading to misclassifications.
- The small dataset may hinder the model's ability to capture nuanced patterns and generalize effectively.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Ayme Arango, Jesus Perez-Martin, and Arniel Labrada. 2022. HateU at SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 581–584, Seattle, United States. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared

Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Shijing Chen, Usman Naseem, Imran Razzak, and Flora Salim. 2024. Unveiling misogyny memes: A multimodal analysis of modality effects on identification. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1864–1871.

Yuyang Chen and Feng Pan. 2022. Multimodal detection of hateful memes by applying a vision-language pre-training model. *PLOS ONE*, 17(9):1–12.

Sneha Chinivar, M. S. Roopa, J. S. Arunalatha, and K. R. Venugopal. 2024. Identification of misogynistic memes using transformer models. In *Proceedings of International Conference on Advanced Communications and Machine Intelligence*, pages 107–116, Singapore. Springer Nature Singapore.

Maibam Debina Devi and Navanath Saharia. 2021. Misogynous text classification using svm and lstm. In *Advanced Computing*, pages 336–348, Singapore. Springer Singapore.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526.

Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Arantza Casillas, Arantza Díaz de Ilarraza, Nerea Ezeiza, Maite Oronoz, Alicia Pérez, and Olatz Perez-de Viñaspre. 2018. Automatic misogyny identification using neural networks. In *IberEval@ SEPLN*, pages 249–254.

Marinella Paciello, Francesca D'Errico, Giorgia Saleri, and Ernestina Lamponi. 2021. Online sexist meme and its effects on moral and emotional processes in social media. *Computers in human behavior*, 116:106655.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Tathagata Raha, Sagar Joshi, and Vasudeva Varma. 2022. IIITH at SemEval-2022 task 5: A comparative study of deep learning models for identifying misogynous memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 673–678, Seattle, United States. Association for Computational Linguistics.

Jason Ravagli and Lorenzo Vaiani. 2022. JRLV at SemEval-2022 task 5: The importance of visual elements for misogyny identification in memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 610–617, Seattle, United States. Association for Computational Linguistics.

Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing Management*, 60(5):103474.

Elena Shushkevich and John Cardiff. 2018. Misogyny detection and classification in english tweets: The experience of the itt team. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:182.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Harshvardhan Srivastava. 2022. Misogynistic meme detection using early fusion model with graph network. *Preprint*, arXiv:2203.16781.

CUET-NLP_MP@DravidianLangTech 2025: A Transformer-Based Approach for Bridging Text and Vision in Misogyny Meme Detection in Dravidian Languages

Md. Mohiuddin, Md Minhazul Kabir
Kawsar Ahmed and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904103, u1904040, u1804017}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Misogyny memes, a form of digital content, reflect societal prejudices by discriminating against women through shaming and stereotyping. In this study, we present a multimodal approach combining Indic-BERT and ViT-base-patch16-224 to address misogyny memes. We explored various Machine Learning, Deep Learning, and Transformer models for unimodal and multimodal classification using provided Tamil and Malayalam meme dataset. Our findings highlight the challenges traditional ML and DL models face in understanding the nuances of Dravidian languages, while emphasizing the importance of transformer models in capturing these complexities. Our multimodal method achieved F1-scores of 77.18% and 84.11% in Tamil and Malayalam, respectively, securing 6^{th} place for both languages among the participants.

1 Introduction

Memes have evolved into a prevalent form of communication in today's digital landscape. These concise pieces of content blend images and text, making them highly engaging to humorously convey complex ideas and emotions. However, the rapid dissemination of memes can also be exploited to spread harmful narratives particularly misogyny. A concerning trend is the frequent sharing of memes that objectify women, promote gender-based violence and propagate harmful stereotypes (Chen et al., 2024a). Further amplifying this issue, nearly 73% of the women surveyed out of over 900 media workers from 125 countries have faced online violence according to a 2022 UNESCO report (Collett et al., 2022). This unchecked rise of misogynistic meme presents a significant threat to social harmony by normalizing misogynistic attitudes and encouraging a toxic culture of gender-based violence. Therefore, effective mitigation strategies are crucial to counter the negative impact of these memes and ensure a more respectful and safer online environment.

Recent research has begun to explore different aspects of memes such as offensive and hate driven content. However, majority of these studies have primarily focused on high resource languages (Singh et al., 2024; Chen et al., 2024b; Fersini et al., 2021) and limited attention is given to low-resource languages like Tamil and Malayalam. In these languages, the detection of misogynistic content in memes remains underexplored, despite significant growth of such content across social media platforms in these regions. The Shared Task on "Misogyny Meme Detection" at DravidianLangTech@NAACL 2025 (Chakravarthi et al., 2024) aims to fill this gap. As part of this shared task, we developed a multimodal system capable of analyzing both textual and visual elements of memes in Tamil and Malayalam to accurately classify them as misogynistic or non-misogynistic. The key contributions of this work are:

- Explored a variety of models (SVM+TF-IDF, CNN, Bi-LSTM, XLM-R, mBERT) for text, (VGG16, VGG19, ResNet50, ViT, Swin) for image, and combinations of these for multimodal analysis to identify an effective method for misogyny meme detection in Tamil and Malayalam.
- Proposed a multimodal model to effectively detect misogynistic memes.

2 Related Work

In recent years, research has focused on improving misogynistic meme detection due to the rise of harmful content online. Addressing the challenge of misogynistic meme detection. Rehman et al. (2025) designed a novel approach combining MANM, GFRM, CFLM to enhance image-text interaction, refine unimodal features and add content

specific elements. This approach outperformed existing methods by 11.87% and 10.82% in F1 score on MAMI and MMHS150K datasets, respectively. Hasan et al. (2024) introduced the Bengali Meme Dataset (AMemD), created to aid in aggression detection in Bengali memes. Among the models tested, the CNN combined with VGG16 achieved the highest F1 score of 0.738. Singh et al. (2024) introduced a dataset of Hindi-English code-mixed misogyny meme detection dataset and investigated different text-only, image-only and multimodal models. For binary classification, their multimodal model, BiT combined with MuRIL BERT, achieved a Macro F1 score of 0.7319. For multilabel classification, their combined BiT and RoBERTa model achieved a Macro F1 score of 0.527. Likewise, Ahsan et al. (2024) also developed a Bengali aggressive meme dataset and utilized the MAF approach which combines the CLIP model for image encoding and Bangla-BERT for text encoding. This method achieved a weighted F1 scroe of 0.742. Zhang and Wang (2022) proposed a multimodal approach using CLIP and UNITER pretrained models, introducing an ensemble method called PBR that achieved top performance in the SemEval-2022 Task 5 with macro F1 scores of 0.834 and 0.731 for sub-task A and B, respectively. By utilizing MMVAE model that integrates text and image embeddings via a VAE for joint multi-task learning. Gu et al. (2022) proposed a method that achieved a macro F1 score of 0.723. Hakimov et al. (2022) presented a multimodal architecture using pretrained CLIP models for both text and image feature extraction, incorporating an LSTM layer for textual context and a fully connected layer for image features. They achieved a macro-F1 score of 0.834. Zhou et al. (2022) tackled the MAMI task at SemEval-2022 using ERNIE-ViL-Large with techniques like biased word masking, image captioning, ensemble learning, and Perspective API, achieving a Macro F1 score of 0.793.

3 Dataset and Task Description

The Shared Task on "Misogyny Meme Detection" (Chakravarthi et al., 2024, 2025) is a multimodal machine learning challenge that aims to classify misogynistic or non-misogynistic memes from social media platforms in Tamil and Malayalam languages. The presented dataset (Ponnusamy et al., 2024) contains images and a CSV file with "image_id", "labels", and "transcriptions" (text), covering both languages. The given dataset is divided into three sets: train, dev and test. The Malayalam dataset is nearly balanced, while the Tamil dataset is imbalanced. The statistics of the dataset are given in Table 1.

Set	Class	Tamil	Malayalam
Train	Misogynistic	285	259
	Non-misogynistic	851	381
Dev	Misogynistic	74	63
	Non-misogynistic	210	97
Test	Misogynistic	89	78
	Non-misogynistic	267	122

Table 1: Dataset distribution for misogyny meme detection

4 System Overview

This section outlines the methodologies and techniques employed to tackle the problem of misogynistic meme detection, encompassing data preprocessing, feature extraction, and the development of a multimodal classification framework integrating textual and visual modalities. The schematic representation of our proposed methodology is illustrated in Figure 1. The complete source code implementation is available on GitHub[1].

4.1 Data Preprocessing

To ensure data consistency and improve model performance, we applied rigorous preprocessing steps tailored to both textual and visual data.

Text Preprocessing: We cleaned the text data by removing emojis, HTML tags, and duplicate entries. Stopwords and punctuation were filtered out, and tokenization was performed using subword-based tokenizers for transformer-based models. Additionally, Term Frequency-Inverse Document Frequency (TF-IDF) (Takenobu, 1994) was employed to extract statistical text features for classical machine learning models.

Image Preprocessing: All images were resized to a uniform dimension and normalized to enhance model stability. Data augmentation techniques, including random flipping, rotation, and contrast adjustment, were applied to improve model generalization.

4.2 Models

To effectively classify memes in Tamil and Malayalam, we employed a combination of traditional

[1] https://github.com/MohiuddinPrantiq/CUET-NLP_MP-DravidianLangTech-NAACL2025

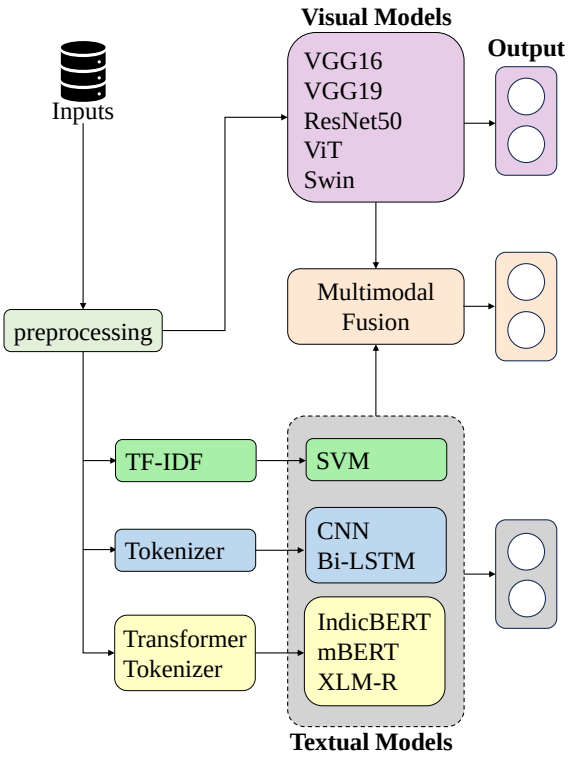


Figure 1: Schematic representation of the proposed system.

machine learning, deep learning, and transformer-based models, leveraging transfer learning to enhance performance without the need for extensive training from scratch (Ibrahim et al., 2020).

4.2.1 Unimodal Models

The provided dataset contains both image and text data, offering an opportunity to investigate individual unimodality models for text and image classification.

Text Classification: Several models were tested for text classification: SVM(Cortes and Vapnik, 1995), CNN(Kim, 2014), Bi-LSTM(Graves and Schmidhuber, 2005), mBERT (Devlin, 2018), and XLM-R (Conneau, 2019) for both Tamil and Malayalam. SVM used a Linear kernel. CNN was implemented with a vocab size of 10,000, an embedding dimension of 100, 100 filters, 0.5 dropout, and filter sizes of 3, 4, and 5. Bi-LSTM used the same vocab size, dropout, and embedding dimension as CNN, with a hidden dimension of 256. Both CNN and Bi-LSTM were trained for 5 epochs with a batch size of 32. mBERT and XLM-R, pretrained transformers from Hugging Face (Wolf, 2020), were trained for 3 epochs, with a batch size of 16 and a learning rate of 2e-5.

Image Classification: For image classification, we used pretrained models: VGG16, VGG19 (Simonyan, 2014), ResNet50 (He et al., 2016), Google's Vision Transformer (Alexey, 2020), and Microsoft's Swin Transformer (Liu et al., 2021) for both Tamil and Malayalam memes. All models were trained for 5 epochs with a learning rate of 1e-4, Adam optimizer, and a batch size of 32.

4.2.2 Multimodal Models

In this task, We need such a system that can analyze both text and images in memes (multimodal) using separate branches for each data type. In our exploration, Different models performed better for each language and modality. For Tamil, SVM and mBERT outperformed other models in text classification, while ResNet50 and Swin performed better in image classification. For Malayalam, SVM and XLM-R were more effective in text, while VGG16 and Swin were superior for images. For multimodal analysis, we selected the top two models from each modality in both languages and concatenated their features to improve prediction accuracy. This resulted in four different combinations for each language. All combinations were trained for 5 epochs with a learning rate of 2e-5.

4.2.3 Proposed Method

Our proposed multimodal fusion model integrates IndicBERT (Kakwani et al., 2020) for textual representation and ViT-Base-Patch16-224 (Wu et al., 2020) for visual feature extraction. The feature embeddings from both modalities are concatenated and passed through a dense classification layer. This fused model was trained for 5 epochs with a batch size of 16 and a learning rate of 2e-5, achieving the best overall performance.

5 Results and Analysis

Table 2 presents a comparative analysis of unimodal and multimodal approaches. In text classification, transformer models outperformed traditional ML and DL models, with mBERT achieving 58.29% in Tamil and XLM-R attaining 73.86% in Malayalam, highlighting their strong contextual understanding. SVM+TF-IDF performed better than CNN and Bi-LSTM, likely due to its efficiency in handling smaller datasets.

For image classification, ResNet50 (68.71%) and VGG16 (80.98%) surpassed vision transformers, suggesting that optimized CNN architectures

Model	Tamil	Malayalam
Text Classification		
SVM+TF-IDF	57.69	65.38
CNN	31.50	50.65
Bi-LSTM	44.84	58.82
XLM-R	23.53	**73.86**
mBERT	**58.29**	56.49
Image Classification		
VGG16	64.05	**80.98**
VGG19	21.57	77.94
ResNet50	**68.71**	75.95
ViT	66.24	73.28
Swin	67.88	80.82
Multimodal Classification		
Tamil		
SVM+ResNet50	45.53	-
SVM+Swin	41.92	-
mBERT+ResNet50	**67.76**	-
mBERT+Swin	64.10	-
Malayalam		
XLM-R+VGG16	-	79.14
XLM-R+Swin	-	**82.58**
SVM+VGG16	-	68.67
SVM+Swin	-	72.73
Proposed Model		
IndicBERT+ViT	**77.18**	**84.11**

Table 2: Comparison of Macro F1-scores across different models on the test set.

remain competitive. However, the close performance of Swin and ViT indicates vision transformers' potential when trained on larger datasets.

In multimodal classification, fusing textual and visual features improved performance. mBERT+ResNet50 reached 67.76% in Tamil, while XLM-R+Swin achieved 82.58% in Malayalam. Our proposed IndicBERT+ViT model obtained the highest F1-scores: 77.18% in Tamil and 84.11% in Malayalam, demonstrating the advantages of multimodal learning. The performance gap between Tamil and Malayalam models suggests that dataset characteristics, including class balance and linguistic complexity, play a key role.

Several key insights emerge from these results. First, text classification performance was lower in Tamil, likely due to class imbalance and the complexity of Tamil script variations. Second, multimodal models consistently outperformed unimodal models, reaffirming the importance of leveraging complementary information sources. However, the lower performance of mBERT+ResNet50 in Tamil compared to ResNet50 alone suggests that feature fusion strategies need refinement. Future work could explore advanced fusion techniques, such as attention-based feature alignment or adaptive weighting, to enhance multimodal learning further. Additionally, dataset balancing techniques may help mitigate bias and improve model generalization, particularly in underrepresented classes.

5.1 Error Analysis

To gain deeper insights into the performance of our proposed model, we conducted a comprehensive error analysis encompassing both quantitative and qualitative evaluations.

Quantitative Analysis

Figure 2 presents the confusion matrices for Tamil and Malayalam meme classification.

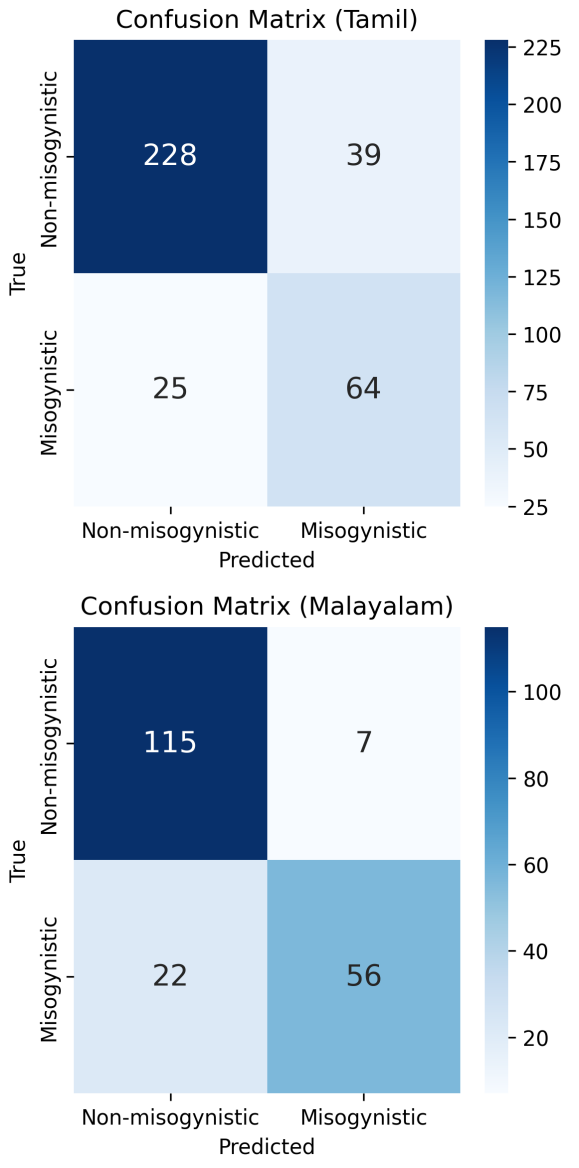


Figure 2: Confusion matrices for Tamil and Malayalam

The model achieved 82.02% accuracy in Tamil (292/356) and 85.5% in Malayalam (171/200), with higher misclassification rates for the positive class (28% vs. 14.5% in Tamil, 5.7% in Malayalam). This disparity stems from dataset imbalance, which limited the model's learning capacity. Notably, the model performed 6.93% better in Malayalam, likely due to a more balanced dataset.

Qualitative Analysis

Figures 3 and 5 highlight correct and incorrect predictions for Tamil, while figures 4 and 6 do the same for Malayalam. While the model accurately classified misogynistic memes in some cases, errors in others suggest an inability to capture nuanced distinctions, primarily due to dataset imbalance. The limited dataset size may have also restricted the model's ability to learn diverse representations, leading to misclassification in complex cases. Furthermore, the simple concatenation-based feature fusion strategy may not have effectively captured intricate relationships between textual and visual modalities, causing the model to overlook subtle cues that differentiate misogynistic from non-misogynistic content. Please see Appendix A for the visualization of the sample memes.

Text Sample	Actual	Predicted
ID_335: causally ignoring when someone complaining about me _vishal_sammy_ EVEL MEWES = = VERA LEVEL	Non-misogynistic	Non-misogynistic
ID_1576: Mano PhotoGrid IRAL Bsru பாத்துக்கிட்டே ருக்கல போல் (Mano is on PhotoGrid IRAL, just watching without participating.)	Non-misogynistic	Misogynistic
ID_1149: OUI MO $ மருமகள் மாமியார் த னக்கு டீ போட தெரியுமா $ 0 டீ போட தெரியுமாவா குக்கர்ல பால ஊத்தி மூணு விசில் விட்டு அப்டியே பொங்கவிடுவோம் (Oh! Does the daughter-in-law know how to make tea for the mother-in-law? Does she know how to make tea or not? Just pour milk into the cooker, give three whistles, and let it boil over like that.)	Misogynistic	Non-misogynistic
ID_1064: எனக்கொரு சினேகிதி சினேகிதி. தென்றல் மாதிரியா..? ல்ல.மெண்டல் மாதிரி. rr (I have a friend, a friend. Like a breeze? No, like a mental one. Haha!)	Misogynistic	Misogynistic

Figure 3: Sample predictions in Tamil.

Text Sample	Actual	Predicted
ID_945: ഇതിന് മാത്രം പിള്ളാരെ ഇവർക്കെങ്ങനെ cilşma??.... (How can these people be blamed for this alone??....)	Non-misogynistic	Non-misogynistic
ID_543: കളിതുടങ്ങി രണ്ടുമിനുട്ട് കഴിഞ്ഞപ്പോൾ ഭർത്താ ഹി ഹി മോളൂ ഭാര്യ പി പി സുനേഷ് എന്നാ വൻമരം വീണു അല്ലെ (Two minutes after the game started, husband: Ha.. ha.. dear... wife: P.P. Suresh, isn't it like a big tree fell?)	Non-misogynistic	Misogynistic
ID_545: അയലത്തെ ചേച്ചിയോട് കളി ചോദിച്ചതിന് ശേഷം നായകൻ ചേച്ചി ഒന്നും പറഞ്ഞില്ല ചേച്ചി കിടക്കു വാ ഞാൻ പോയി അടച്ചിട്ട് വരാം (After asking the neighbor's sister for a game, the hero said, 'Sister didn't say anything... Sister is resting. I'll go, close the door, and come back.)	Misogynistic	Non-misogynistic
ID_61: എന്തൊക്കെ ആകിയിട്ടും ഒരു മെന വരുന്നില്ലലോ Nikhi.. (Whatever happens, isn't there any sign of a message, Nikhi)	Misogynistic	Misogynistic

Figure 4: Sample predictions in Malayalam.

6 Conclusion

This study examined the performance of unimodal and multimodal approaches using various machine learning, deep learning, and transformer models for misogynistic meme detection in Tamil and Malayalam. The results show that the combination of Indic-BERT and ViT achieved the highest F1-scores of 77.18% and 84.11% in Tamil and Malayalam, respectively. These findings highlight the effectiveness of transformer models in both unimodal and multimodal classification for this task. To address the current constraints, future work can explore improved fine-tuning strategies, leverage large language models (LLMs) and alternative transformer architectures, and adopt more sophisticated feature fusion techniques such as attention mechanisms and cross-modal transformers. Additionally, integrating Vision-Language Models (VLMs) such as CLIP (Radford et al., 2021) and LXMERT (Tan and Bansal, 2019) could enhance multimodal understanding by jointly learning from textual and visual features in a more coherent manner. Further efforts in handling data imbalance—including data augmentation, reweighting strategies, SMOTE, class-weighted loss, and adversarial training along with expanding dataset size and diversity or incorporating a CNN-based backbone with transfer learning could also improve classification accuracy and robustness.

Limitations

Despite the strong performance of our proposed approach, several limitations remain. One key challenge is misclassification, likely influenced by dataset imbalance, which affects generalization. Additionally, our use of IndicBERT, which supports 12 Indian languages, may not extend effectively to languages outside its training scope. Another constraint arises from our straightforward feature fusion strategy, which relies on simple concatenation and may not fully capture the complex interactions between textual and visual modalities.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500.

Dosovitskiy Alexey. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Shijing Chen, Usman Naseem, Imran Razzak, and Flora Salim. 2024a. Unveiling misogyny memes: A multimodal analysis of modality effects on identification. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1864–1871.

Shijing Chen, Usman Naseem, Imran Razzak, and Flora D. Salim. 2024b. Unveiling misogyny memes: A multimodal analysis of modality effects on identification. *Companion Proceedings of the ACM on Web Conference 2024*.

Clementine Collett, Livia Gouvea Gomes, Gina Neff, et al. 2022. *The effects of AI on the working lives of women*. UNESCO Publishing.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elisabetta Fersini, Giuliano Rizzi, Aurora Saibene, and Francesca Gasparini. 2021. Misogynous meme recognition: A preliminary study. In *International Conference of the Italian Association for Artificial Intelligence*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Yimeng Gu, Ignacio Castro, and Gareth Tyson. 2022. Mmvae at semeval-2022 task 5: A multi-modal multi-task vae on misogynous meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 700–710.

Sherzod Hakimov, Gullal S Cheema, and Ralph Ewerth. 2022. Tib-va at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. *arXiv preprint arXiv:2204.06299*.

Md Hasan, Shawly Ahsan, Moshiul Hoque, and M. Dewan. 2024. *MuLAD: Multimodal Aggression Detection from Social Media Memes Exploiting Visual and Textual Features*, pages 107–123.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2020. AlexU-BackTranslation-TL at SemEval-2020 task 12: Improving offensive language detection using data augmentation and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1881–1890, Barcelona (online). International Committee for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated

dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Mohammad Zia Ur Rehman, Sufyaan Zahoor, Areeb Manzoor, Musharaf Maqbool, and Nagendra Kumar. 2025. A context-aware attention and graph neural network-based multimodal framework for misogyny detection. *Information Processing & Management*, 62(1):103895.

Karen Simonyan. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Aakash Singh, Deepawali Sharma, and Vivek Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Thomas Wolf. 2020. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision. *Preprint*, arXiv:2006.03677.

Jing Zhang and Yujin Wang. 2022. SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596, Seattle, United States. Association for Computational Linguistics.

Ziming Zhou, Han Zhao, Jingjing Dong, Ning Ding, Xiaolong Liu, and Kangli Zhang. 2022. DD-TIG at SemEval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 563–570, Seattle, United States. Association for Computational Linguistics.

A Appendix

True: Non-misogynistic | Image ID: 335 | Predicted: Non-misogynistic

True: Non-misogynistic | Image ID: 954 | Predicted: Non-misogynistic

True: Non-misogynistic | Image ID: 1576 | Predicted: Misogynistic

True: Non-misogynistic | Image ID: 543 | Predicted: Misogynistic

True: Misogynistic | Image ID: 1149 | Predicted: Non-misogynistic

True: Misogynistic | Image ID: 545 | Predicted: Non-misogynistic

True: Misogynistic | Image ID: 1064 | Predicted: Misogynistic

True: Misogynistic | Image ID: 61 | Predicted: Misogynistic

Figure 5: Sample memes in Tamil

Figure 6: Sample memes in Malayalam

CUET_NetworkSociety@DravidianLangTech 2025: A Transformer-based Approach for Detecting AI-Generated Product Reviews in Low-Resource Dravidian Languages

Sabik Aftahee*, **Tofayel Ahmmed Babu***, **MD Musa Kalimullah Ratul***
Jawad Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904005, u1904024, u1904071, u1704039}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

E-commerce platforms face growing challenges regarding both consumer trust and review authenticity because of the growing number of AI-generated product reviews. Low-resource languages such as Tamil and Malayalam face limited investigation by AI detection techniques because these languages experience constraints from sparse data sources and complex linguistic structures. The research team at CUET_NetworkSociety took part in the AI-Generated Review Detection contest during the DravidianLangTech@NAACL 2025 event to fill this knowledge void. Using a combination of machine learning, deep learning, and transformer-based models, we detected AI-generated and human-written reviews in both Tamil and Malayalam. Among the approaches used, DistilBERT was found to be better suited to detect AI-Generated Reviews, which underwent an advanced preprocessing pipeline and hyperparameter optimization using the Transformers library. This approach achieved a Macro F1-score of 0.81 for Tamil (Subtask 1), securing 18^{th} place, and a score of 0.72 for Malayalam (Subtask 2), ranking 25^{th}.

1 Introduction

Online authenticity and reliability face serious obstacles because of the recent growth of AI-generated content in the current era. Product reviews experience direct negative impacts because customers heavily depend on them during purchasing decisions. These reviews are being generated by AI, often mimicking human reviews. This rise in AI-generated reviews has far-reaching implications, as it undermines trust in online marketplaces, misleads consumers, and distorts market dynamics (Raja et al., 2023). Thus, the need to detect those contents is very imminent. Researchers created solid detection methods for AI-generated content in different languages through advanced deployments of natural language processing and machine learning systems (LekshmiAmmal et al., 2022). However, most studies have focused mainly on high-resource languages such as English and Spanish, leaving low-resource languages such as Tamil and Malayalam underrepresented (Hegde and Shashirekha, 2021). The Dravidian language family poses distinctive detection challenges because of its complex morphological structure combined with semantic richness and various dialectal variations (Coelho et al., 2023). Research on AI-generated review detection in Tamil and Malayalam languages faces negligible attention despite their economic relevance due to the limited available datasets and the intricate linguistic structures of these languages (Krishnan et al., 2024). Online review credibility and user trust in the specified regions become essential to remedy. The research establishes a reliable method for detecting product evaluations created by AI in Tamil and Malayalam by focusing on this specific problem. The system explored various machine learning (ML), deep learning (DL), and transformer-based models to overcome the linguistic challenges of detecting AI-generated product reviews in Dravidian languages. The critical contributions of this work are:

- Investigated several ML, DL, and transformer-based models to detect AI-generated reviews in Tamil and Malayalam.
- Evaluated the performance of employed models and provided a comparative analysis to identify the most effective approach for detecting AI-generated content in these Dravidian languages.

2 Related Work

Recent research on fake review detection has focused on using machine learning (ML) and deep learning (DL) techniques to tackle this problem

*Authors contributed equally to this work.

in various datasets. Barbado et al. (2019) proposed the Fake Feature Framework (F3) to detect fake reviews using user-centric and review-centric features, working with a custom Yelp consumer electronics dataset and the DOSA dataset, achieving an F1-score of 82% with AdaBoost. Raheem and Chong (2024) compared deep learning models (LSTM, CNN, hybrid) with transformers (DistilBERT) for fake review detection, utilizing the Yelp Reviews dataset, and found DistilBERT achieved 96% accuracy. Abd-Alhalem et al. (2024) integrated deep learning with aspect-based sentiment analysis using the OSF dataset, achieving 97.73% accuracy with an LSTM-based model. Vashist et al. (2024a) employed ensemble machine learning techniques (XGBoost, Random Forest) combined with BERT for detecting fake reviews on the OSFHOME dataset, achieving 98.2% accuracy with BERT. Deshai and Bhaskara Rao (2023) explored hybrid deep learning models (CNN-LSTM and LSTM-RNN) with GloVe embeddings for fake review and rating detection on Amazon Unlocked Mobile and Hotel datasets, reaching 93.07% accuracy. Ennaouri and Zellou (2023) reviewed various ML techniques and ensemble voting for fake review detection, reporting 97.5% accuracy on the CloudArmor dataset. Saini and Khatarkar (2023) analyzed fake news detection methods, which can be applied to fake reviews, using the WELFake dataset, achieving 96.73% accuracy. Veda et al. (2024) proposed a hybrid model combining BERT embeddings and ensemble methods (Random Forest, XGBoost) on the Public Fake Reviews dataset, achieving 86.45% accuracy with a stacking classifier. Rajesh et al. (2023) utilized sentiment analysis with traditional ML classifiers on Amazon Reviews, achieving 85% accuracy with Logistic Regression and Count Vectorizer. Wagh et al. (2024) applied Random Forest and NLP techniques on the Amazon Yelp Academic dataset to detect spam reviews, achieving 89.49% accuracy. Vashist et al. (2024b) used CNN and SVM models for detecting fraudulent reviews on a custom dataset, with CNN achieving 89% accuracy. V et al. (2023) provided a general overview of ML techniques for fake review detection but did not specify a dataset or accuracy. Alkomah and Sheldon (2023) reviewed advancements in fake news detection techniques, which could be adapted for fake review detection, but did not provide specific performance metrics. Sharma et al. (2023) explored hybrid deep learning models, integrating Bi-LSTM and CNN for fake review detection, highlighting the effectiveness of combining contextual and sequential information, with the highest accuracy reported at 95%. Transformers have revolutionized AI-generated content detection with models like BERT, RoBERTa, and their multilingual variants. LekshmiAmmal et al. (2022) demonstrated the potential of transformers in detecting toxic spans in Tamil, while Bafna et al. (2023) developed a RoBERTa-BiLSTM hybrid model that achieved a significant boost in accuracy for AI-generated text detection. Moreover, Coelho et al. (2023) focused on Malayalam, showcasing the efficacy of TF-IDF combined with ensemble ML models for detecting fake reviews in low-resource languages, achieving a macro F1 score of 0.831.

3 Task and Dataset Descriptions

For the goal of detecting AI-generated product reviews in Tamil and Malayalam languages, we utilized datasets specifically curated for this task (Premjith et al., 2025). The datasets consist of training, validation, and test data with detailed distributions as outlined below.

3.1 Tamil Dataset

The Tamil data set comprises a balanced distribution of human-generated and AI-generated reviews. Table 1 presents the statistics of the Tamil dataset in the training, validation, and test sets.

Classes	**Train**	**Test**	W_t	U_w
AI	405	48	3583	1423
Human	403	52	2428	1281
Total	808	100	6011	2704

Table 1: Class-wise distribution of training and test sets for Tamil, where W_T denotes the total number of words, and U_W denotes the number of unique words

3.2 Malayalam Dataset

The Malayalam dataset also maintains a balanced distribution across AI-generated and human-written reviews. Detailed statistics are shown in Appendix A. Both datasets provide a nearly equal class distribution across AI-generated and human-written reviews, ensuring a balanced evaluation setup. The unique word counts highlight the linguistic diversity and vocabulary range in both Tamil and Malayalam datasets. The implementa-

tion details are available at the link[1].

4 Methodology

The following section details a complete set of procedures along with methodologies which tackle the existing text classification hurdles described earlier. Figure 1 illustrates the abstract process of detecting AI-generated reviews. Our method uses machine learning alongside deep learning and transformer-based models while optimizing and tuning these models to boost their performance in text classification applications.

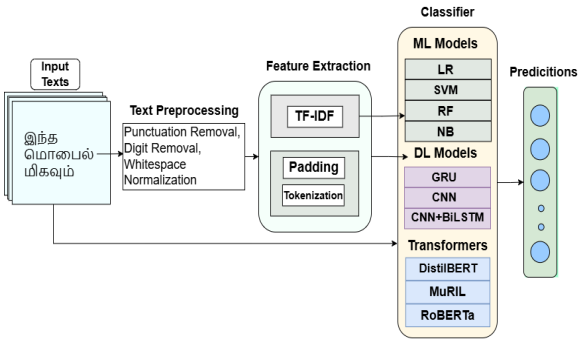


Figure 1: Abstract process of AI Generated Review detection

4.1 Pre-processing and Feature Extraction

We used extensive pre-processing techniques to normalize input data which created essential conditions for model training success. The preprocessing step involved removal of tags, punctuation, and numbers to ensure uniformity. Additionally, the text data was transformed using TF-IDF vectorization, which has been shown to effectively identify important words within a document.

4.2 ML Models

We used baseline machine learning models such as LR, SVM, RF, and Naive Bayes for the first assessments. Multiple metrics including accuracy, precision, recall, and F1 score were used for evaluation to measure model performance. TF-IDF vectorization was implemented which transformed text data into its top 1000 terms for efficient classification. Logistic regression received configuration adjustments for better convergence performance by setting its maximum iteration threshold to 1000.

[1] https://github.com/pr0ximaCent/DravidianLangtech-2025

4.3 DL Models

Additionally, a combination of deep learning architectures such as CNNs and BiLSTM networks were used to further assess the work. These models are particularly better at capturing complex patterns in text data, making them suitable for the nuanced demands of natural language understanding. The training of these models was systematically conducted, involving the tuning of hyperparameters like the number of layers, dropout rates, and learning rates to optimize performance. These models utilized embeddings of dimension 128 and were optimized with the Adam optimizer at a learning rate of 1e-3, training on batches of 32 samples. The classification output was derived using a sigmoid activation function.

4.4 Transformer Models

The highlight of our methodology was the application of transformer-based models to detect AI-generated product reviews in Dravidian languages, celebrated for their efficiency and robustness in handling various NLP tasks (Fariello et al., 2024). We fine-tuned three transformer-based models (MuRIL-BERT, RoBERTa and DistilBERT) on our dataset, which entailed several pivotal steps: text data tokenization using the transformers library tokenizer, integration of early stopping, and dynamic learning rate adjustments to forestall overfitting while expediting convergence to an optimal model state. Training was meticulously executed using Hugging Face's `Trainer` API, incorporating strategies such as batch size optimization and validation-based tuning to ensure the model's effectiveness (Forte and Marotta, 2024; Raja and Wani, 2023). The models were trained for up to four epochs, with periodic evaluations to adjust training parameters based on real-time performance metrics. Each model was validated with a distinct set to ensure better reliability & generalization on unseen data (Chaka, 2024; Ara et al., 2024).

5 Result Analysis

We observed the ability of Machine Learning (ML) and Deep Learning (DL) with Transformer-based models to identify AI-generated write-ups from authentic human reviews within Tamil and Malayalam datasets. The measurement of classifier performance included precision (P), recall (R), F1-score (F1), and accuracy (A). A comprehensive summary of model performance is presented in

Table 2.

Classifier	P	R	F1	A
Malayalam				
Logistic Regression (LR)	0.58	0.60	0.59	0.61
SVM	0.55	0.56	0.555	0.57
Random Forest (RF)	0.53	0.53	0.53	0.54
Naive Bayes (NB)	0.50	0.48	0.49	0.51
CNN	0.63	0.64	0.635	0.65
GRU	0.66	0.66	0.66	0.67
CNN-LSTM	0.65	0.65	0.65	0.66
CNN-BiLSTM	0.67	0.68	0.67	0.69
MuRIL-BERT	0.68	0.68	0.68	0.69
RoBERTa	0.67	0.67	0.67	0.68
DistilBERT	**0.75**	**0.71**	**0.72**	**0.75**
Tamil				
Logistic Regression (LR)	0.60	0.62	0.61	0.63
SVM	0.57	0.58	0.57	0.59
Random Forest (RF)	0.55	0.55	0.55	0.56
Naive Bayes (NB)	0.52	0.50	0.51	0.53
CNN	0.65	0.66	0.655	0.67
GRU	0.68	0.68	0.68	0.69
CNN-LSTM	0.67	0.67	0.67	0.68
CNN-BiLSTM	0.69	0.70	0.69	0.71
MuRIL-BERT	0.70	0.70	0.70	0.71
RoBERTa	0.69	0.69	0.69	0.70
DistilBERT	**0.85**	**0.78**	**0.81**	**0.76**

Table 2: Performance Comparison of Classifiers Across ML, DL, and Transformer Models for Tamil and Malayalam

Logistic Regression (LR) achieved F1 scores of 0.59 for Malayalam and 0.61 for Tamil, highlighting its effectiveness in modeling linear relationships. However, traditional machine learning models like SVM, Random Forest (RF), and Naive Bayes (NB) performed weaker, with F1 scores between 0.49 and 0.56, showing limitations in capturing complex semantic patterns in text. These models rely heavily on manual feature engineering and may struggle with high-dimensional data like text. Their inability to automatically capture semantic and syntactic complexities resulted in a subpar performance in language tasks.

The CNN-BiLSTM model achieved F1 scores of 0.67 for Malayalam and 0.69 for Tamil, effectively capturing complex relationships and long-range dependencies. Alternative deep learning models like GRU and CNN-LSTM performed competitively, with F1 scores between 0.66 and 0.68. Deep learning architectures can learn hierarchical representations of data, capturing local and global text dependencies. This ability enabled them to understand context and semantics better than traditional models, leading to improved performance.

Transformer-based models, especially DistilBERT, outperformed traditional ML and deep learning models, achieving the highest F1 scores (0.75 for Malayalam, 0.81 for Tamil) using dynamic contextual embeddings. MuRIL-BERT and RoBERTa followed closely with F1 scores between 0.68 and 0.70, showcasing strong performance, particularly for low-resource languages. Traditional ML methods and CNN-BiLSTM demonstrated adequate performance, but transformers surpassed them to achieve better results. Transformers utilize self-attention mechanisms to effectively weigh the importance of different words in a sentence. This capability enabled them to capture complex patterns and contextual relationships more efficiently than previous architectures, leading to superior performance in language understanding tasks. Hyperparameter tunings have been discussed in Appendix B.

6 Error Analysis

An in-depth error analysis was conducted using both quantitative and qualitative methods to evaluate the performance of the proposed model.

6.1 Quantitative Error Analysis

To further understand the performance of the models, a quantitative analysis was performed using confusion matrices for Tamil and Malayalam. Figures 2 and 3 illustrate the confusion matrices for both languages.

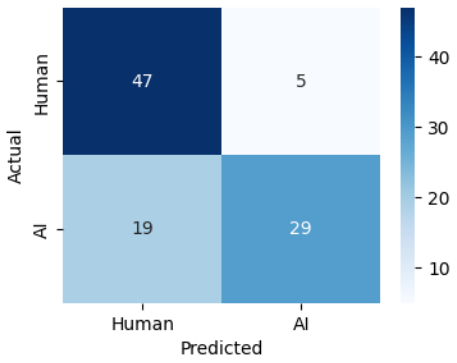


Figure 2: Confusion Matrix for Tamil Dataset

The Tamil confusion matrix shows 47 correctly identified human reviews, with 5 misclassified as AI. However, 19 AI reviews were wrongly labeled as human, while 29 were correctly classified. Similarly, in Malayalam, 79 human reviews were correctly identified, but 21 were misclassified. The model also correctly predicted 74 AI reviews, though 26 were mistaken for human, highlighting challenges in AI text detection.

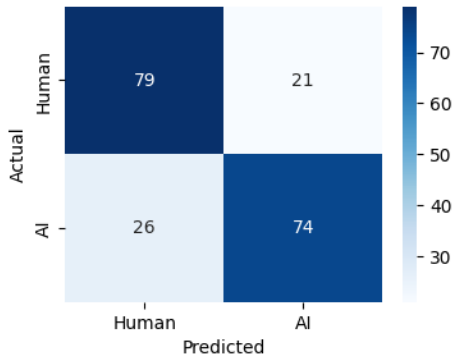


Figure 3: Confusion Matrix for Malayalam Dataset

Text Sample	Actual	Predicted
അടച്ച പൈസ നഷ്ടപ്പെടാതെ കിട്ടണമെങ്കിൽ എൽ.ഐ.സി മാത്രേ ഒള്ളൂ. 100% സോവെറിക്ൻ ഗുയാരണ്ടീ.	Human	Human
കോവയ്ക്ക ഉപ്പിലിട്ത് ഞാൻ ഇതുവരെ കഴിച്ചിട്ടില്ല, കഴിക്കാൻ മനസില്ല.	AI	Human
ഞാൻ 19227 മന്തിലി ടാറ്റ അടക്കുന്നു ഫോർ 5 വർഷം. മെച്യൂരിറ്റി ര്ം വർഷം . ഏകദേശം 10 കോടിക്ക് മുകളിൽ റിട്ടേൺ കിട്ടും	Human	Human
പഴകിയ മീനും കറികളും കഴിച്ചു പല പ്രാവശ്യം ഫുഡ് പൊയ്സണിംഗിനെ നേരിട്ടിട്ടുണ്ട്.	AI	Human
കോവയ്ക്ക ഉപ്പിലിട്ത് ഞാൻ ഇതുവരെ കഴിച്ചിട്ടില്ല, കഴിക്കാൻ മനസില്ല.	AI	AI

Figure 5: Few examples of predicted outputs by the proposed (DistilBERT) model for Malayalam

Some errors in quantitative analysis, as reflected in confusion matrices, stem from overlapping linguistic patterns between AI-generated and human-written text. The model struggles to differentiate when AI text mimics human fluency or human reviews exhibit generic phrasing, leading to misclassifications.

6.2 Qualitative Error Analysis

To complement the quantitative analysis, a qualitative examination of the misclassified examples was conducted. Figures 4 and 5 present a few representative examples of predicted outputs by the model.

Text Sample	Actual	Predicted
என் ஹெட்செட்/ஈர்பாட் பிராண்ட் பயன்படுத்தியபோது, சில நேரங்களில் ஆடியோ குவாலிட்டி சரியில்லாமல், சத்தம் மிகவும் குறைந்துவிடுகிறது.	AI	AI
நான் அண்மையில் வாங்கிய ஒரு குக்கர் ஆரோக்கிய உணவு தயாரிப்பதற்காக சரியான தேர்வாக இல்லை.	AI	Human
அணிவதற்கு நன்றாக இருக்கும்	Human	Human
அதிக வாசனை வாந்தி	Human	AI
அதிகமக பயன்பதுதினல் தலை சுட்ரும்	Human	Human

Figure 4: A few examples of predicted outputs by the proposed (DistilBERT) model for Tamil.

The qualitative analysis revealed misclassifications where AI reviews were labeled as human-written due to colloquial language, and human reviews were mistaken for AI-generated due to generic phrasing. These errors highlight the need for improved contextual differentiation and feature extraction to reduce misclassifications.

7 Conclusion

This research explored the detection of AI-generated product reviews in low-resource Dravidian languages, specifically Tamil and Malayalam, using machine learning, deep learning, and transformer-based models. The study found that traditional ML models like Logistic Regression and SVM struggled to capture the intricate linguistic features of these languages. Deep learning approaches, such as CNN-BiLSTM, improved performance by better modeling text dependencies. However, the transformer-based DistilBERT model demonstrated the highest effectiveness, achieving the best F1-scores for both Tamil and Malayalam datasets. The research outcome confirms that transformer models demonstrate high capability when used for text classification in languages with minimal resources. The next step should concentrate on using extensive datasets as well as better fine-tuning methods and contextual elements to enhance the accuracy rate.

Limitations

Despite the contributions in detecting AI-generated reviews in Tamil and Malayalam, several limitations remain: (i) Pre-trained transformer models like DistilBERT may be limited by their training corpus, affecting their ability to capture the nuances of these languages. (ii) The small datasets used constrained the models' generalization to unseen data. (iii) Linguistic complexities, such as code-mixing and dialect variations, present challenges in accurate text classification.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Samia M. Abd-Alhalem, Hesham A. Ali, Naglaa F. Soliman, Abeer D. Algarni, and Hanaa Salem Marie. 2024. Advancing e-commerce authenticity: A novel fusion approach based on deep learning and aspect features for detecting false reviews. *IEEE Access*, 12:1–17.

Bushra Alkomah and Frederick Sheldon. 2023. Advancements in fake news detection using machine and deep learning models: Comprehensive literature review. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1–6.

Anjuman Ara, Md Sajadul Alam, Kamrujjaman, and Afia Farjana Mifa. 2024. A comparative review of ai-generated image detection across social media platforms. *Global Mainstream Journal of Innovation, Engineering Emerging Technology*, 3(1):11–19.

R. Bafna, M. Jain, and P. Sharma. 2023. Roberta and bilstm hybrid architecture for ai-generated text detection. In *Proceedings of the 2023 International Conference on Natural Language Processing*, pages 233–241.

Rodrigo Barbado, Oscar Araque, and Carlos A. Iglesias. 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing Management*, 56(4):1234–1244.

C. Chaka. 2024. Differentiating between ai-generated and human-written text using ai detection tools. *Journal of Applied Learning and Teaching*, 7(1):118–126.

Sharal Coelho, Asha Hegde, and Hosahalli Lakshmaiah Shashirekha. 2023. Mucs@dravidianlangtech2023: Malayalam fake news detection using machine learning approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 288–292.

N Deshai and B Bhaskara Rao. 2023. Deep learning hybrid approaches to detect fake reviews and ratings. *Journal of Scientific & Industrial Research*, 82:120–127.

Mohammed Ennaouri and Ahmed Zellou. 2023. Machine learning approaches for fake reviews detection: A systematic literature review. *Journal of Web Engineering*, 22:821–848.

Serena Fariello, Giuseppe Fenza, Flavia Forte, Mariacristina Gallo, and Martina Marotta. 2024. Distinguishing human from machine: A review of advances and challenges in ai-generated text detection. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(5):1–12.

F. Forte and M. Marotta. 2024. Machine learning models for ai-generated text analysis. *Computational Intelligence Review*, 14(3):234–249.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2021. Urdu fake news detection using ensemble of machine learning models. *CEUR Workshop Proceedings*, pages 132–141.

S. Krishnan, R. Babu, and P. Nair. 2024. Detecting ai-generated text: A study on the performance of ml and dl models in dravidian languages. *Journal of Artificial Intelligence Research*, 17(4):254–271.

Hariharan LekshmiAmmal, Manikandan Ravikiran, and Anand Kumar Madasamy. 2022. Nitk-it nlp@tamilnlp-acl2022: Transformer based model for toxic span identification in tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 75–78, Dublin, Ireland. Association for Computational Linguistics.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, Sajeetha Thavareesan, and Prasanna Kumar Kumaresan. 2025. Overview of the shared task on detecting ai generated product reviews in dravidian languages: Dravidianlangtech@naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Mafas Raheem and Yi Chien Chong. 2024. E-commerce fake reviews detection using lstm with word2vec embedding. *CIT.2024*, 100:70–80.

K. Raja and S. Wani. 2023. Multilingual sentiment analysis for fake review detection. *International Journal of Computational Linguistics*, 12(1):88–102.

R. Raja, A. Kumar, and S. Joseph. 2023. Fake news detection in low-resource languages: Challenges and advancements. *Computational Linguistics Review*, 15(2):123–137.

N Rajesh, AC Ramachandra, Ayush Tomar, Heman Kumawat, Anurag Prasad, and Ramprasad Poojary. 2023. Fake reviews detection based on sentiment analysis using ml classifiers. *IEEE International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*.

Parul Saini and Virendra Khatarkar. 2023. A review on fake news detection using machine learning. *SMART MOVES JOURNAL IDSCIENCE*, 9:6–9.

A. Sharma, S. Kumar, and R. Gupta. 2023. Evaluating convolutional neural networks for text classification tasks in low-resource languages. *Journal of Computational Linguistics*, 49(2):123–135.

Arpitha S V, Ashwitha H N Jois, Bhargavi V M, Deeksha A H, and Sreedevi S. 2023. Detecting fake reviews in e-commerce platform. *International Journal of Advanced Research in Computer and Communication Engineering*, 12:1–6.

Ansh Vashist, Arul Keswani, Varda Pareek, and Tarun Jain. 2024a. Detecting fake reviews on e-commerce platforms using machine learning. In *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, pages 1–6.

Ansh Vashist, Arul Keswani, Varda Pareek, and Tarun Jain. 2024b. Detecting fraudulent reviews in e-commerce platforms. *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, pages 1–6.

Chitti Reddy Veda, Muni Sekhar Velpuru, Namburu Apoorva, Hammikolla Akshaya, N Vishnu, and Sai Prakhyath Siripuram. 2024. Fake review identification using hybrid fusion of machine learning and natural language processing techniques. *IEEE Access*.

Yogansh Wagh, Sarfaraz Ali, Apurva Bobade, Aditi Bhalekar, and Rajaram Ambole. 2024. E-commerce spam review detection using machine learning. *Journal of Information Systems and Renewable Energy Management (JISREM)*, pages 1–4.

A Class-wise Distribution of Malayalam Dataset

Table A.1 shows class-wise distribution of training and test sets for the Malayalam language, including the number of total and unique words in each category. The dataset is divided into AI-generated and Human-written texts, with an equal split between training and test samples. The statistics, such as the total words (W_T) and unique words (U_W), highlight the lexical diversity within each class. This information is crucial for understanding the dataset composition and its impact on model training and evaluation.

Classes	**Train**	**Test**	W_t	U_w
AI	400	100	5174	3138
Human	400	100	8201	4819
Total	800	200	13375	7957

Table A.1: Class-wise distribution of training and test sets for Malayalam where W_T and U_W, denotes total and unique words respectively .

B Tuned Hyperparameters

Table B.1 shows the fine-tuned hyperparameters for AI vs. Human text classification tasks using DistilBERT. A learning rate of 5×10^{-5} ensures stable convergence, while a batch size of 16 balances memory and training stability. The model trains for 4 epochs with a max sequence length of 256 to capture longer texts efficiently. Cross-entropy loss is used for classification, with AdamW as the optimizer and a weight decay of 0.01 to prevent overfitting. Gradient accumulation steps (2) simulate a larger batch size of 32, while 300 warmup steps and a linear learning rate scheduler help stabilize training. These hyperparameters were fine-tuned to maximize accuracy while maintaining computational efficiency.

Hyperparameter	**Value**
Learning Rate	5×10^{-5}
Per Device Batch Size	16
Number of Epochs	4
Max Sequence Length	256
Loss Function	Cross-Entropy Loss
Optimizer	AdamW
Weight Decay	0.01
Gradient Accumulation Steps	2
Warmup Steps	300
Learning Rate Scheduler	Linear

Table B.1: Tuned hyperparameters used for the AI vs. Human text classification task using DistilBERT.

CUET_NetworkSociety@DravidianLangTech 2025: A Multimodal Framework to Detect Misogyny Meme in Dravidian Languages

MD Musa Kalimullah Ratul*, **Sabik Aftahee***, **Tofayel Ahmmed Babu***
Jawad Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904071, u1904024, u1904005, u1704039}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

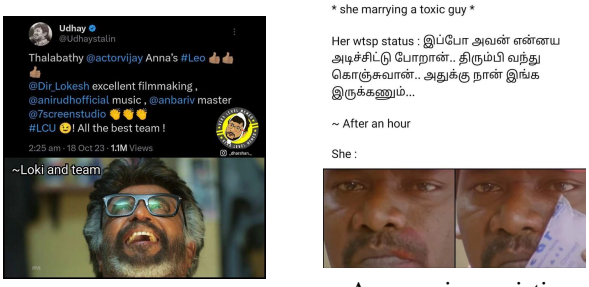
Abstract

Memes are widely used to communicate on social networks. Some memes can be misogynistic, conveying harmful messages to individuals. Detecting misogynistic memes has become a broader challenge, particularly for low-resource languages like Tamil and Malayalam, due to their linguistic morphology. This paper presents a multi-modal deep learning approach for detecting misogynistic memes in Tamil and Malayalam. The proposed model integrates fine-tuned ResNet18 for extracting visual features and `ai4bharat/indic-bert` for analyzing textual content. The fusion model was then applied to make predictions on the test dataset. The model achieved a macro F1 score of 76.32% for Tamil and 80.35% for Malayalam. Our technique helped secure 7th and 12th positions for Tamil and Malayalam, respectively.

1 Introduction

In the past few years, the popularity of the social media platform has gained a huge response from individuals, where a multi-modal phenomenon called meme has been introduced to us. The meme is generally an image with some contextual texts of that image. Apart from humorous contents, memes also can carry harmful messages, such as misogyny, which is the hatred of, contempt for, or prejudice against women or girls. The satirical nature of such content makes identifying misogynistic memes particularly challenging, as they often employ nuanced, context-dependent signals that can evade straightforward detection (Rizzi et al., 2023). To illustrate this distinction, Figure 1 compares a misogynistic meme and a non-misogynistic meme, highlighting the subtle yet significant differences in their messaging.

Misogynistic memes pose a threat to society that basically try to normalize hatred against women and gender bias. Existing research has explored multimodal approaches to identifying misogyny contents in memes with both textual and visual features to achieve higher accuracy. For instance, Rizzi et al. (Rizzi et al., 2023) investigated 4 unimodal and 3 multimodal approaches to determine which source of information contributes more



A misogynistic meme

A non-misogynistic meme

Figure 1: Examples of memes: (a) Misogynistic and (b) Non-misogynistic.

to the detection of misogynous memes. Similarly, Singh et al. (Singh et al., 2024) investigated misogyny detection in Hindi-English code-mixed memes, showcasing the complexities of handling low-resource languages and multimodal content. The shared task on multitask meme classification further emphasized the importance of distinguishing misogynistic and trolling behaviors, providing valuable datasets and benchmarks for advancing the field (Chakravarthi et al., 2024).

Even with this much advancement, research on misogynistic memes in low-resource languages such as Tamil and Malayalam remains very much challenging. Our paper introduces a multi-modal deep learning framework for detecting misogynistic memes in Tamil and Malayalam.

The key contributions of this work include:

- Proposed a multimodal framework designed to detect misogynistic memes while capturing the linguistic and cultural nuances of Tamil and Malayalam.
- Investigated the performance of several ML, DL, and transformer-based models for misogynistic meme detection, highlighting challenges in low-resource languages through quantitative and qualitative error analysis.

2 Related Work

Multimodal approaches, which leverage the interplay between text and images, have shown promising advancements. Gasparini et al. (Gasparini et al., 2022) tackled the automatic detection of misogynistic con-

*Authors contributed equally to this work.

tent in memes using multimodal data, evaluating a dataset of 800 memes (400 misogynistic and 400 non-misogynistic). Singh et al. (Singh et al., 2024) introduced *MIMIC*, a dataset containing 5,054 Hindi-English code-mixed memes for misogyny detection, demonstrating the effectiveness of multimodal fusion, where ViT+RoBERTa achieved a macro f1 score of 0.7532 for Multi-label Misogyny Classification. Rizzi et al. (Rizzi et al., 2023) addressed biases in misogynistic meme detection by proposing a debiasing framework on an 800-meme dataset, achieving up to a 61.43% improvement in prediction probabilities. Here Visual-BERT achieved a macro f1 score of 0.84. Chakravarthi et al. (Chakravarthi et al., 2024) organized a shared task on multitask meme classification for misogyny and trolling in Tamil and Malayalam memes, achieving macro F1 scores of 0.73 for Tamil and 0.87 for Malayalam. Ponnusamy et al. (Ponnusamy et al., 2024) introduced the *MDMD* dataset for misogyny detection in Tamil and Malayalam memes. Hegde et al. (Hegde et al., 2021) focused on Tamil troll meme classification, demonstrating the effectiveness of attention-based transformers. They achieve an overall F1-score of 0.96 by using images for classification using ViT.

Hossain et al. (Hossain et al., 2024) proposed an *Align before Attend* strategy for multimodal hateful content detection on the MUTE (Bengali code-mixed) and MultiOFF (English) datasets, achieving F1-scores of 69.7% and 70.3%, respectively. Ahsan et al. (Ahsan et al., 2024) developed the *MIMOSA* dataset for target-aware aggression detection in Bengali memes, introducing a multimodal aggression fusion (MAF) model that outperformed state-of-the-art approaches. Here the ViT achieved the highest weighted F1-score of 0.582 and for the textual-only Bangla-BERT model surpassed all unimodal models with a weighted F1-score of 0.641. Rahman et al. (Rahman et al., 2024) also proposed a comprehensive multimodal approach for abusive language detection in Tamil, incorporating textual, acoustic, and visual features. The study utilized a dataset annotated for abusive and non-abusive classes, employing models such as ConvLSTM, 3D-CNN, and BiLSTM. Their weighted late fusion model, ConvLSTM+BiLSTM+MNB, achieved the highest macro F1 score of 71.43%. Conneau et al. (Conneau et al., 2020) developed XLM-R, a cross-lingual representation learning model trained on CommonCrawl data in 100 languages, achieving 85.0% accuracy on XNLI. Feng et al. (Feng et al., 2022) introduced LaBSE, a language-agnostic BERT model for multilingual sentence embeddings, enhancing cross-lingual understanding. Here LaBSE achieved a highest accuracy of 95.3%

Tan and Le (Tan and Le, 2020) proposed EfficientNet for image recognition, achieving 84.4% top-1 accuracy on ImageNet, making it a strong candidate for feature extraction in multimodal tasks. He et al. (He et al., 2016) introduced deep residual learning with ResNet, achieving a 3.57% top-1 error rate on ImageNet, widely adopted for image feature extraction in multimodal studies. Zhou et al. (Zhou et al., 2015) proposed C-LSTM for text classification, leveraging CNN-LSTM hybrid models on various text classification datasets, though specific accuracy results were not mentioned. Their implementation C-LSTM got the highest accuracy of 94.6%. Arevalo et al. (Arevalo et al., 2017) proposed gated multimodal units for information fusion, focusing on multimodal data representation. Here GMU achieved a macro-f1 score of 0.541.

These works collectively underscore the potential of multimodal strategies in detecting misogyny and other forms of online abuse. They also highlight the challenges, including handling implicit content, managing noisy or low-resource data, and ensuring model fairness and generalizability across diverse cultural contexts.

3 Task and Dataset Description

The task focuses on developing models for detecting misogynistic content in memes from social media. These models analyze both textual and visual components to classify memes as *Misogynistic* or *Non-Misogynistic*. Misogynistic content includes text or visuals targeting women with harmful, offensive, or derogatory intent, while non-misogynistic memes align with respectful communication standards. This task is part of the DravidianLangTech@NAACL 2025 Shared Task on Misogyny Meme Detection (Chakravarthi et al., 2025a,b). Additionally, prior shared tasks have explored related issues, such as the LT-EDI@EACL 2024 shared task on multitask meme classification, which examined misogynistic and troll memes in Tamil and Malayalam (Chakravarthi et al., 2024). The dataset consists of Tamil and Malayalam memes, including text extracted from images, corresponding visual data, and labels. This dataset is based on the work by (Ponnusamy et al., 2024), who introduced the MDMD (Misogyny Detection Meme Dataset) to address the propagation of misogyny, gender-based bias, and harmful stereotypes in online memes. Tables 1 and 2 show the dataset distribution.

Dataset	Misogynistic	Non-Misogynistic	Total
Train	259	381	640
Dev	63	97	160
Test	78	122	200

Table 1: Dataset distribution for Malayalam memes

Dataset	Misogynistic	Non-Misogynistic	Total
Train	285	851	1136
Dev	74	210	284
Test	89	267	356

Table 2: Dataset distribution for Tamil memes

The Malayalam dataset contains 640 training, 160 development, and 200 test samples (Table 1). Similarly, the Tamil dataset comprises 1136 training, 284 development, and 356 test samples (Table 2). Both datasets

include text extracted from memes, visual data, and classification labels, supporting multimodal approaches for effective classification. The GitHub link[1] contains both the source code and datasets.

4 Methodology

The proposed approach for misogyny meme detection leverages Multimodal deep learning by integrating textual and visual information extracted from memes. The process extracts of image features using Convolutional neural networks (CNNs) and the transformation of text data using transformer-based language models. The classification models are trained independently for each modality before employing a fusion strategy to improve the overall classification performance. Figure 2 presents the schematic framework for the Multimodal misogyny meme detection system.

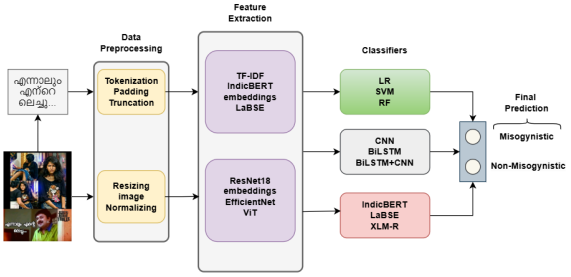


Figure 2: Schematic framework for Misogyny Meme Detection

4.1 Visual Approach

For the image modality, feature extraction was performed using two state-of-the-art pre-trained convolutional neural networks (CNNs): ResNet18 (He et al., 2016) and EfficientNet-B4 (Tan and Le, 2020). These architectures were selected due to their high performance in visual recognition tasks and their ability to capture hierarchical spatial features. Each meme image was resized to a standard resolution of 224×224 pixels and normalized to maintain consistent pixel value distributions. Random horizontal flipping, rotation, and brightness adjustments were applied to improve model generalization. The extracted deep features from the CNN architectures were passed through fully connected layers before being used for classification. The ResNet18 model, with its residual connections, helps mitigate vanishing gradient issues and captures robust spatial information. On the other hand, EfficientNet-B4 leverages compound scaling to balance network depth, width, and resolution, optimizing performance while maintaining efficiency. These extracted features were further combined in multimodal fusion strategies for improved classification performance.

[1] https://github.com/MusaRatul/Misogyny-Meme-Detection-in-Dravidian-Languages

4.2 Textual Approach

The textual component of the meme data was processed using advanced transformer-based models, which have demonstrated state-of-the-art performance in various natural language processing tasks (Vaswani et al., 2017). The textual features were extracted using IndicBERT (Kunchukuttan et al., 2020), LaBSE (Feng et al., 2022), and XLM-RoBERTa (Conneau et al., 2020). IndicBERT, specifically designed for low-resource Indian languages, was employed for its ability to handle complex linguistic variations in Tamil and Malayalam. LaBSE, a multilingual model, captured robust sentence-level embeddings, making it suitable for cross-lingual tasks. XLM-RoBERTa, a cross-lingual transformer model, provided strong contextual embeddings for diverse languages, including Tamil and Malayalam. The preprocessing pipeline included tokenization, punctuation removal, and stop-word filtering. The extracted embeddings from these models were passed to classification layers where classical machine learning classifiers such as Logistic Regression, Support Vector Machines (SVM), and Random Forests were explored.

4.3 DL Approach

Beyond traditional classifiers, deep learning architectures were employed to enhance textual classification. CNN, BiLSTM, and a hybrid BiLSTM+CNN model were tested for their ability to capture and model complex textual patterns. The CNN model captured local semantic patterns using convolutional filters (Kim, 2014), while BiLSTM processed sequential dependencies in textual data using bidirectional long short-term memory networks. For the BiLSTM model, contextual embeddings extracted from transformer-based models IndicBERT, LaBSE, and XLM-RoBERTa were input representations. These embeddings provided rich semantic text representations, enabling the BiLSTM to model temporal dependencies while effectively preserving contextual meaning. The hybrid BiLSTM+CNN model combined the local feature extraction capability of CNNs with the sequential modeling power of BiLSTMs, further enhancing feature extraction for textual data (Zhou et al., 2015). All deep learning models were trained with categorical cross-entropy loss and optimized using the Adam optimizer with a learning rate of 0.001 and a batch size of 32.

4.4 Multimodal Fusion Approach

To integrate textual and visual features, two fusion techniques were explored. In the late fusion approach, independent models for text and image classification were trained separately, and their predictions were combined using weighted averaging. Various weight allocations were tested, with uniform weighting yielding the best results. The concatenation-based fusion approach merged the text and image features at an intermediate layer before the final classification. This allowed the model to learn joint representations of multimodal data, enhanc-

ing discriminatory power. Late fusion was preferred over early fusion to preserve modality-specific feature representations, prevent interference between text and image embeddings, and leverage the strengths of pre-trained models. The findings align with previous studies, such as Arevalo et al. (Arevalo et al., 2017) and Kiela et al. (Kiela et al., 2020), which highlight the benefits of combining image and text modalities for robust multimodal learning. The proposed multimodal approach demonstrated significant improvements in misogyny meme detection, leveraging both the hierarchical spatial representations from CNN models and the contextual embeddings from transformer-based language models.

5 Results Analysis

The classification performance is summarized in Tables 3 and 4.

Model	P	R	F1
LR	0.8000	0.7000	0.7250
SVM	0.7400	0.6900	0.7050
RF	**0.8550**	0.6150	0.6300
CNN	0.3750	0.5000	0.4286
BiLSTM	0.7837	0.7416	0.7582
BiLSTM+CNN	0.7415	0.7397	0.7406
LaBSE+EfficientNet-B4	0.7687	0.7360	0.7494
ViT+XLM-R	0.8240	0.7453	0.7620
IndicBERT+ResNet18	0.8095	**0.7772**	**0.7632**

Table 3: Performance Comparison on Tamil Dataset

Model	P	R	F1
LR	0.7276	0.7314	0.7292
SVM	0.7373	0.7373	0.7373
RF	0.6843	0.6686	0.6725
CNN	0.3050	0.5000	0.3789
BiLSTM	0.8077	0.7998	0.8032
BiLSTM+CNN	0.8030	0.7934	0.7973
LaBSE+EfficientNet-B4	0.7844	0.7857	0.7851
ViT+XLM-R	0.8115	0.8085	0.7999
IndicBERT+ResNet18	**0.8329**	**0.8162**	**0.8035**

Table 4: Performance Comparison on Malayalam Dataset

For the Tamil dataset (Table 3), the IndicBERT+ResNet18 model achieved the highest performance, with a Precision (P) of 0.8095, Recall (R) of 0.7772, and an F1-score of 0.7632. Among other models, ViT+XLM-R also performed well with an F1-score of 0.7620, followed by BiLSTM (0.7582) and LaBSE+EfficientNet-B4 (0.7494). Traditional machine learning models such as Logistic Regression (LR) and Support Vector Machines (SVM) lagged behind, with F1-scores of 0.7250 and 0.7050, respectively. CNN exhibited the lowest performance, with an F1-score of 0.4286, highlighting its limitations in capturing complex text-visual relationships. The superior performance of IndicBERT+ResNet18 suggests that ResNet18 extracted spatially rich visual features that complemented the textual representations from IndicBERT better than ViT, which may not have captured fine-grained spatial details as effectively.

For the Malayalam dataset (Table 4), IndicBERT+ResNet18 again outperformed all models, achieving a Precision of 0.8329, Recall of 0.8162, and an F1-score of 0.8035. Close contenders included BiLSTM (0.8032), ViT+XLM-R (0.7999), and BiLSTM+CNN (0.7973), showing that deep learning models performed consistently well. Traditional machine learning approaches like LR and SVM had moderate performance, with F1-scores of 0.7292 and 0.7373, respectively. CNN showed the weakest performance, with an F1-score of 0.3789, reinforcing its inefficacy in handling the multimodal nature of the task. This outcome suggests that ResNet18 provided more structured and discriminative visual embeddings than ViT in this context, leading to a more potent synergy with IndicBERT for multimodal learning.

6 Error Analysis

Both quantitative and qualitative error analyses were conducted to gain deeper insights into the performance of the proposed model.

6.1 Quantitative Analysis

To further understand the performance of the models, a quantitative analysis was performed using confusion matrices for Tamil and Malayalam datasets. Figures 3 and 4 illustrate the confusion matrices for both languages.

The confusion matrices reveal that the model performs well in identifying explicit misogynistic content, correctly classifying a high proportion of such memes. However, there are notable errors in detecting non-misogynistic memes and implicit misogyny, leading to false positives and false negatives. This indicates that while the model effectively captures explicit cues, it struggles with subtler textual or contextual features.

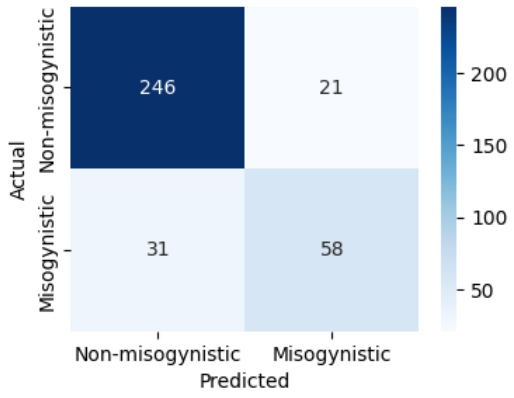


Figure 3: Confusion matrix for the top-performing model (IndicBERT+ResNet18) in Tamil misogynistic meme detection

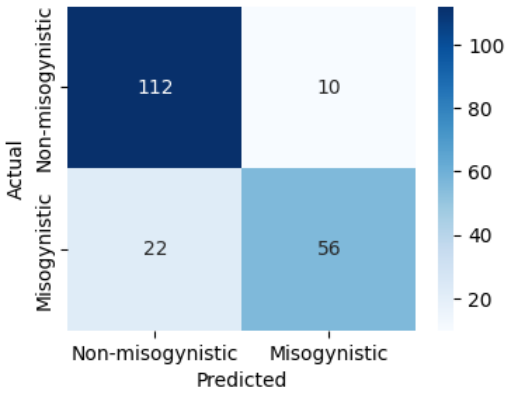


Figure 4: Confusion matrix for the top-performing model (IndicBERT+ResNet18) in Malayalam misogynistic meme detection

6.2 Qualitative Analysis

To complement the quantitative analysis, a qualitative examination of the misclassified examples was conducted. Tables 5 and 6 present representative examples of predicted outputs by the IndicBERT+ResNet18 model for Tamil and Malayalam datasets.

Image	Actual	Predicted
	Misogyny	Misogyny
	Misogyny	Non-Misogyny

Table 5: Examples of predicted outputs from the IndicBERT+ResNet18 model for Tamil misogynistic meme detection

The qualitative analysis reveals that several misclassifications occurred due to linguistic subtleties, sarcasm, and neutral visual elements. Some Tamil memes contained subtle expressions or indirect language that the model failed to interpret correctly. Malayalam memes often employed sarcasm or implicit misogynistic references, making it challenging for the model to capture the intended meaning. In cases where visual elements were neutral or did not reinforce textual cues, the model struggled to combine features effectively, leading to misclassification. These findings suggest that future work should focus on improving the fusion strategy and incorporating external knowledge to better capture contextual and implicit cues, thereby reducing such misclassifications.

Image	Actual	Predicted
	Non-Misogyny	Misogyny
	Non-Misogyny	Non-Misogyny

Table 6: Examples of predicted outputs from the IndicBERT+ResNet18 model for Malayalam misogynistic meme detection

7 Conclusion

The proposed multimodal approach effectively combines textual and visual features for misogyny detection in Tamil and Malayalam memes, achieving competitive results. By integrating TF-IDF, IndicBERT embeddings, ResNet18, and EfficientNet with machine learning, deep learning, and transformer-based models, this paper demonstrates the potential of multimodal fusion in tackling complex classification tasks. Future work will focus on improving sarcasm detection, as sarcasm often overlaps with misogynistic content and remains challenging to identify accurately. Additionally, there is significant potential to enhance performance by leveraging larger multilingual datasets that include more diverse and representative samples across different languages and cultural contexts.

Limitations

- The model struggles with detecting implicit cues in meme text and visuals, leading to occasional misclassification of sarcastic misogynistic content.
- IndicBERT's limitations affect performance for Tamil and Malayalam, particularly in handling nuanced language structures, impacting overall classification accuracy.
- Memes with complex multimodal sarcasm require improved fusion strategies, necessitating future research on enhanced dataset augmentation and fine-tuned multilingual transformers for better linguistic and contextual understanding.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M. Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500, St. Julian's, Malta. Association for Computational Linguistics.

John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. 2017. Gated multimodal units for information fusion. *Preprint*, arXiv:1702.01992.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025a. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025b. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in Brief*, 44:108526.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Uvce-iiitt@dravidianlangtech-eacl2021: Tamil troll meme classification: You need to pay more attention. *Preprint*, arXiv:2104.09081.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah Masud Preum. 2024. Align before attend: Aligning visual and textual features for multimodal hateful content detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 162–174, St. Julian's, Malta. Association for Computational Linguistics.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020. Supervised multimodal bitransformers for classifying images and text. *Preprint*, arXiv:1909.02950.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N. C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *Preprint*, arXiv:2005.00085.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Md. Rahman, Abu Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das, and Mohammed Moshiul Hoque. 2024. Binary_Beasts@DravidianLangTech-EACL 2024: Multimodal abusive language detection in Tamil based on integrated approach of machine learning and deep learning techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 212–217, St. Julian's, Malta. Association for Computational Linguistics.

Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing Management*, 60(5):103474.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Mingxing Tan and Quoc V. Le. 2020. Efficientnet: Rethinking model scaling for convolutional neural networks. *Preprint*, arXiv:1905.11946.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A c-lstm neural network for text classification. *Preprint*, arXiv:1511.08630.

CUET_NetworkSociety@DravidianLangTech 2025: A Transformer-Driven Approach to Political Sentiment Analysis in Tamil X (Twitter) Comments

Tofayel Ahmmed Babu*, **Sabik Aftahee***, **MD Musa Kalimullah Ratul***
Jawad Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904005, u1904024, u1904071, u1704039}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Social media has become an established medium of public communication and opinions on every aspect of life, but especially politics. This has resulted in a growing need for tools that can process the large amount of unstructured data that is produced on these platforms providing actionable insights in domains such as social trends and political opinion. Low-resource languages like Tamil present challenges due to limited tools and annotated data, highlighting the need for NLP focus on understudied languages. To address this, a shared task has been organized by DravidianLangTech@NAACL 2025 for political sentiment analysis for low-resource languages, with a specific focus on Tamil. In this task, we have explored several machine learning methods such as SVM, AdaBoost, GB, deep learning methods including CNN, LSTM, GRU BiLSTM, and the ensemble of different deep learning models, and transformer-based methods including mBERT, T5, XLM-R. The mBERT model performed best by achieving a macro F1 score of 0.2178 and placing our team 22nd in the rank list.

1 Introduction

Understanding and interpreting human emotions and opinions expressed in text has become a vital aspect of natural language processing (NLP), particularly in the context of social media and public discourse. With the arrival of social media networks such as X (Twitter), the need for advanced sentiment analysis tools, especially politically, is once more highlighted. Political sentiment analysis can provide key insights into the opinion of the population, party identification, and social concerns which are of great interest to policymakers, analysts, and political operators. However, it is difficult to find an advanced sentiment analysis tool for any low-resource languages such as Tamil (Chen et al., 2015). Tamil is a speech of more than 80 million people worldwide (Jain et al., 2020), and therefore a source of a huge corpus of internet dialogue, especially social media. However, Tamil is poorly represented in NLP literature due to the morphological complexity and lack of annotated datasets. Efforts to create annotated datasets for Tamil sentiment analysis (Chakravarthi, 2020), along with tools for morphological analysis (Sarveswaran et al., 2021), have provided a solid foundation for further advancements in the field. Additionally, the development of hybrid architectures that combine deep learning techniques (Ramesh Babu, 2022) with multilingual transformer models (Roy and Kumar, 2021) has shown considerable promise in addressing the challenges posed by the complex linguistic nature of Tamil. However, despite these strides, the lack of substantial recent advancements in comprehensive toolkits for Tamil NLP continues to hinder progress, especially in specialized tasks such as political sentiment analysis. This paper aims to build on these foundational efforts by proposing an improved system for political sentiment analysis in Tamil tweets. Our contributions include:

- Developed a transformer-based system for political multiclass sentiment analysis of Tamil X (Twitter) comments.
- Investigated various machine learning, deep learning, and transformer-based models for Tamil political sentiment analysis and conducted an in-depth error analysis to evaluate the performance of these models.

2 Related Work

Recent advancements in sentiment analysis for low-resource languages have increasingly relied on multilingual transformers such as XLM-R (Conneau, 2019) and IndicBERT (Kannan et al., 2021), which have demonstrated effectiveness in Tamil

*Authors contributed equally to this work.

sentiment classification. Integrating contextualized embeddings (Kenton and Toutanova, 2019; Liu, 2019) and transfer learning techniques (Ruder et al., 2019) has further addressed the challenges of data scarcity. Additionally, CNN-based architectures (Kim, 2014) continue to contribute to feature extraction, reinforcing the role of deep learning in sentiment analysis for low-resource languages. Nazir et al. (2025) proposed a transformer-based approach using CMSA-mBERT for multiclass sentiment analysis (Positive, Negative, and Neutral) on the CMDSA-24 dataset, achieving F1 score of 79.87% and the result shows huge improvements over traditional methods. Khan et al. (2025) proposed an attention-based, stacked CNN-BiLSTM model for Urdu sentiment analysis, improving feature extraction and sequential pattern recognition. Evaluated on UCSA-21 and UCSA datasets, it achieved an accuracy of 83.12% and 78.91%, respectively.

For Tamil-specific NLP tasks, Sarveswaran et al. (2021) created tools specifically for Tamil morphology, providing a baseline for further research. Jain et al. (2020) developed a high-quality Tamil-to-English translation system that outperforms Google Translator (which might indirectly be useful for sentiment analysis problems) and its effects on text representation. Attai et al. (2024) provided a useful insight into political discourse and analysis. They used machine learning techniques (SVM, RF, XGBoost) to analyze public sentiment in the 2023 Nigerian General Elections and it was established that 43% of the tweets were Neutral, 33% Positive, and 24% Negative, with XGBoost achieving the highest accuracy of 93%. Sampath and Supriya (2024) explored sentiment analysis on code-mixed data using translation-based preprocessing and transformer models, achieving 94% accuracy with DistilBERT for Tamil-English and 92% for Hindi-English. Results highlight the effectiveness of specialized NLP models over traditional translation tools. Moreover, a study by K et al. (2023) on textual sentiment analysis in Tamil and Tulu code-mixed texts employed SVM and ensemble models with fastText and TF-IDF, achieving F1-scores of 0.14 (Tamil) and 0.20 (Tulu).

3 Task and Dataset Description

In the shared task, the provided dataset (Chakravarthi et al., 2025) comprises three CSV files for training, validation, and testing, containing 4352, 544, and 544 data points, respectively. Notably, the dataset exhibits class imbalance across its sentiment categories. The dataset includes seven sentiment classes: *Opinionated, Sarcastic, Neutral, Positive, Substantiated, Negative,* and *None of the Above*. Detailed class-wise statistics are presented in Table 1.

Class	Train	Val	Test	W_T	UW_T
Opinionated	1361	153	171	31748	13540
Sarcastic	790	115	106	17231	8717
Neutral	687	84	70	13975	7075
Positive	575	69	75	13251	6459
Substantiated	412	52	51	10310	5679
Negative	406	51	47	9079	4997
None of the Above	171	20	25	1619	1193
Total	**4352**	**544**	**544**	**97213**	**47660**

Table 1: Class-wise distribution of train, validation, and test set for political multiclass sentiment analysis of Tamil X (Twitter) comments, where val, W_T, and UW_T denote validation, total words in each class, and total unique words in each class, respectively

For enhanced visualization, bar charts have been included in the Appedix B. The implementation details of the tasks will be found in the GitHub repository[1].

4 Methodology

Various machine learning (ML), deep learning (DL), and transformer-based models were utilized to create a strong baseline, as shown in Figure 1.

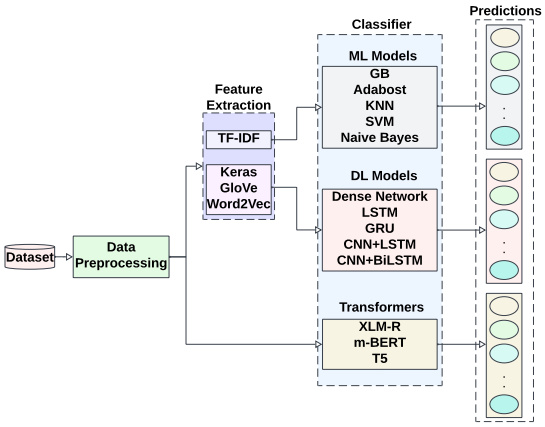


Figure 1: Schematic process of political multiclass sentiment analysis of Tamil X (Twitter) comments

4.1 Data Preprocessing

The dataset contained 4352, 544, and 544 samples for the training, validation, and testing sets, re-

[1] https://github.com/5pace4/NAACL-2025

spectively. To ensure data integrity, instances with missing values in the content or labels columns were removed. Preprocessing was primarily aimed at standardization with the use of the unidecode library that aimed to remove the accents, lower the text to set it to default, and clean up the non-alphanumerical and digital forms with regular expression, removing non-alphabet texts and numbers and conversion to lowercase. Other preprocessing steps included clearing double or more spaces into one and paragraph formatting. Then sentiment labels are converted to numerical values using *LabelEncoder* so that the machine learning model can handle data efficiently.

4.2 Feature Extraction

Feature extraction in NLP transforms raw text into machine-readable numerical representation through various techniques. It varied for machine learning, deep learning, and transformer-based models in this paper. Machine learning models (SVM, GB, AdaBoost, KNN, Naive Bayes) used TF-IDF vectorization with unigram and bigram features, mapping 5,440 samples (4,352 training, 544 validation, 544 test) into 5,000-dimensional matrices, yielding shapes of (4,352 × 5,000) for training and (544 × 5,000) for validation and test, respectively. Deep learning models utilized TensorFlow Keras Tokenizer for tokenization and pre-trained embeddings like word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) (glove.6B.300d.txt), forming a (5,001 × 300) embedding matrix, including an OOV token. Transformer models (mBERT, XLM-R, T5) employed contextualized word embeddings with sub-word tokenization, dynamically adjusting representations for sentiment classification. This multilayered feature extraction enhances model efficiency and accuracy.

4.3 Machine Learning Models

Various machine learning algorithms were used to classify sentiment in Tamil such as SVM, GB, AdaBoost, and KNN. SVM finds a separate hyperplane, while GB reduces bias variation, and AdaBoost improves performance by iterating on weak learners. KNN performs the classification of samples based on their proximity in feature space. All models were trained using features extracted with TF-IDF, providing a strong baseline for the classification of Tamil sentiment.

4.4 Deep Learning Models

The proposed approach applies deep learning methods to model both local and global textual features for the sentiment classification of Tamil X (Twitter) comments. This study tries many variants of architecture such as Dense Networks, LSTM, GRU, CNN+LSTM, and CNN+BiLSTM. The Dense Network with Batch Normalization served as the baseline model, integrating fully connected layers with dropout and batch normalization to mitigate overfitting. LSTM and BiLSTM captured long-term dependencies in text sequences, while CNN extracted local patterns using convolutional filters. Additionally, pre-trained embeddings (GloVe and Word2Vec) were explored, serving as input layers for LSTM, GRU, and CNN architectures, which were fine-tuned during training. All models were trained on tokenized and padded text sequences, ensuring uniform input dimensions across different architectures. Table 2 provides the hyperparameters used for deep learning models in sentiment classification, including LSTM, GRU, CNN+LSTM, and CNN+BiLSTM.

Model	Embedding	Layers/Units	Epochs	Batch	Opt.
Dense Net	None	Dense: 256, 128	15	64	Adam
LSTM	GloVe	LSTM: 128, 64	20	32	Adam
GRU	Word2Vec	GRU: 128, 64	15	32	Adam
CNN+LSTM	GloVe	Conv1D: 128, LSTM: 64	15	32	Adam
CNN+LSTM	Word2Vec	Conv1D: 128, LSTM: 64	15	32	Adam
CNN+BiLSTM	Word2Vec	Conv1D: 128, BiLSTM: 64	15	32	Adam
CNN+BiLSTM	GloVe	CNN: 256, BiLSTM: 128, 64	30	32	Adam
GRU	GloVe	GRU: 128, 64	15	32	Adam

Table 2: Hyperparameters of deep learning for sentiment classification

The models use either GloVe or Word2Vec embeddings and are trained with the Adam optimizer. Batch sizes range from 32 to 64, while epochs vary between 15 and 30, ensuring a balance between computational efficiency and model performance.

4.5 Transformer-Based Models

Transformer-based models including mBERT (Devlin, 2018), XLM-R (Conneau, 2019), and T5 (Ni et al., 2021) were fine-tuned for sentiment classification in Tamil text. All these models are based on pre-trained contextual embeddings that capture subtle semantic nuances. Of these, the best macro F1 score is obtained by mBERT, benefiting from its robust multilingual training. The cross-lingual task-optimized XLM-R gave competitive results, while T5 effectively handled structured outputs using a sequence-to-sequence approach. Fine-tuning involved task-specific adaptations and optimizing cross-entropy loss. Each model has been tuned

based on a set of several hyperparameters summarised in Table A.1 Appendix A.

5 Result Analysis

This paper introduced the political sentiment classification of Tamil text. The performance of sentiment classification models was evaluated using precision, recall, and macro F1 score across ML, DL, and transformer-based approaches. A summarizing table of various models performance is shown in Table 3. Among the machine learning models, SVM achieved the best F1 score, which is 0.2167. The strength of the SVM lies in its ability to find a separating hyperplane between sentiment classes in a high-dimensional space. AdaBoost got a slightly lower F1 score of 0.2088, while Gradient Boosting (GB) achieved F1 score of 0.2028. However, KNN and NB demonstrated limited performance, achieving F1 scores of 0.1430 and 0.1045, respectively, due to their simplicity and inability to model complex patterns in the dataset effectively.

Classifiers	Precision	Recall	F1 Score
SVM	0.3345	0.2547	0.2167
GB	0.3476	0.2620	0.2028
AdaBoost	0.2288	0.2581	0.2088
KNN	0.2339	0.2203	0.1430
Naive Bayes	0.4360	0.1592	0.1045
Deep Learning Models			
Dense Network	0.1968	0.1637	0.1270
LSTM (G)	0.2636	0.1946	0.1917
CNN+LSTM (G)	0.2743	0.1863	0.1840
CNN+LSTM (W)	0.3032	0.1971	0.1887
CNN+BiLSTM (G)	0.3151	0.2160	0.2166
CNN+BiLSTM (W)	0.3339	0.1991	0.1891
GRU (G)	0.2903	0.1876	0.1868
GRU (W)	0.2671	0.1822	0.1600
Transformer-Based Models			
XLM-R	0.1704	0.1752	0.1317
T5	0.3071	0.2073	0.1937
mBERT	**0.2557**	**0.3190**	**0.2178**

Table 3: Performance comparison of classifiers across ML, DL, and transformer models where G and W denote GloVe and Word2Vec embedding, respectively

In the DL domain, CNN+BiLSTM with GloVe embeddings yielded the highest F1 score of 0.2166, while CNN+BiLSTM with Word2Vec achieved 0.1891. A simpler Dense network had the lowest F1 score of 0.1270, showcasing the limitations of shallow architectures for this task. Other architectures, such as CNN+LSTM with Glove, CNN+LSTM with Word2Vec, GRU with GloVe, GRU with Word2Vec, and LSTM with GloVe achieved F1 scores of 0.1840, 0.1887, 0.1868, 0.1600, and 0.1917, respectively.

The best performance of the transformer-based models was that of mBERT, pre-trained on multilingual datasets with dynamic contextual embeddings, which really worked for the Tamil text, with F1 score of 0.2178 and placed the team 22nd in the final rank list. T5 follows next because of its sequence-to-sequence learning approach, with F1 score of 0.1937. XLM-R obtained the worst F1 score, reaching just 0.1317 probably for being more cross-lingual transfer-focused than fine-tuned for sentiment classification in some languages. Overall, the results show that hybrid deep learning models and transformer-based approaches are immensely better in capturing the rich semantics of Tamil sentiment compared to traditional machine learning methods. However, the modest macro F1 scores across all methods highlight the challenges posed by low-resource languages like Tamil, such as data scarcity and morphological complexity.

6 Error Analysis

Both quantitative and qualitative error analyses were conducted to gain deeper insights into the performance of the proposed model.

6.1 Quantitative Analysis:

The best-performing models were used to conduct a quantitative error analysis, utilizing confusion matrices shown in Figure 2.

The confusion matrix describes performance and challenges concerning seven classes of sentiments in Tamil i.e. *Negative, Neutral, None of the Above, Opinionated, Positive, Sarcastic,* and *Substantiated.* It is noticed that the model has performed well in identifying the *Opinionated* class, followed by *None of the Above* and *Sarcastic*, with the highest number of correct predictions 129 instances, 21, and 19, respectively. However, there were significant misclassification patterns, especially for the *Negative* class, which was almost entirely misclassified, mostly as *Opinionated* with 36 instances. The *Neutral* class was quite confused with *Opinionated,* with 52 instances, showing that there is some difficulty in distinguishing these classes. Though the *Positive* class had 9 correct predictions, a large number of instances were misclassified as *Opinionated,* showing that there is some overlap in semantic features. Moreover, the *Substantiated* class also suffered, with zero correct predictions and heavy misclassifications into

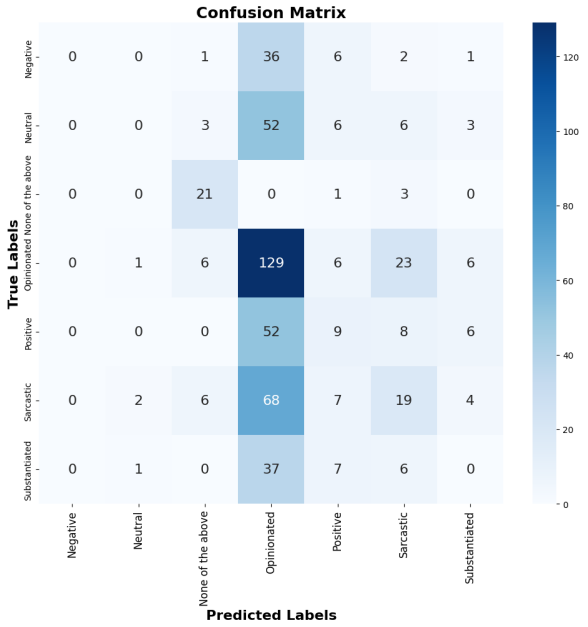


Figure 2: Confusion matrix of the proposed model (fine-tuned mBERT) for political multiclass sentiment analysis

Opinionated and *Positive,* reflecting ambiguities in contextual cues. A major source of these errors seems to be the class imbalance in the dataset. *Opinionated* class dominating others and such imbalance prohibits the model from learning fine-grained features of minority classes, leading to over-reliance on more frequent categories.

Sample Text	Actual Label	Predicted Label
Sample 1: இஸ்லாமிய சகோதரர்களுடன் ரமலான் கொண்டாடிய அதிமுக வேட்பாளர் ராயபுரம் மனோ #royapurammano #admk #chennai #electioncampaign	Neutral	Opinionated
Sample 2: ஓபிஎஸ் - எடப்பாடி போட்டா போட்டி! திடீரென பணிகளை முடுக்கியுள்ள எடப்பாடி! #AIADMK #OPS #EPS #Annamalai #Edappadi #OPanneerselvam #OPSvsEPS #DMK #BJP #Seeman #NaamTamilar #MKStalin #IPL #IPL2023 #CSK #ChennaiSuperKings #RCBVSLSG #ViratvsGambhir #ViratKohli	Negative	Opinionated
Sample 3: mony kathir அரியாத ஜனங்கள்	None of The Above	None of The Above
Sample 4: நன்றி அண்ணா. #மக்களின்_சின்னம்_மைக்	Opinionated	Opinionated

Figure 3: Few examples of predicted outputs by the proposed method (mBERT) for political multiclass sentiment analysis

6.2 Qualitative Analysis:

Figure 3 illustrates the predicted outputs of the proposed model for Tamil political multiclass sentiment analysis based on sample inputs. The model correctly classified Samples 3 and 4 but misclassified *Neutral* (Sample 1) and *Negative* (Sample 2) as *Opinionated,* likely due to contextual bias from hashtags and data imbalance.

7 Conclusion

This study evaluated multiple approaches for classifying political multiclass sentiment in X (Twitter) comments, including ML, DL, and transformer-based models. Among them, mBERT achieved the best macro F1 score (0.2178), benefiting from multilingual pretraining and dynamic contextual embeddings. Hybrid deep learning models, such as CNN+BiLSTM with GloVe, also performed competitively, effectively capturing both local and sequential features. In contrast, traditional ML models struggled with the task's complexity. The overall low F1 scores highlight the challenges of sentiment analysis in a morphologically rich, low-resource language like Tamil. In future work, we aim to mitigate class imbalance using resampling techniques and weighted loss functions, conduct ablation studies to analyze mBERT's performance and explore improvements through data augmentation, fine-tuning, and Tamil-specific preprocessing strategies.

Limitations

Despite the contributions of the current work on political multiclass sentiment analysis of Tamil X (Twitter) comments has several drawbacks. i) As the proposed approach relies on pre-trained transformer-based model, its performance may degrade in scenarios where the context significantly deviates from the data on which the model was originally trained. ii) The focus on Tamil-specific sentiment analysis limits the applicability of the models to other low-resource languages without significant adaptation. iii) The dataset used is imbalanced, which may have impacted the model's ability to generalize across all sentiment categories.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

K Attai, D Asuquo, KE Okonny, AB Johnson, A John, I Bardi, C Iroanwusi, and O Michael. 2024. Sentiment analysis of twitter discourse on the 2023 nigerian general elections. *European Journal of Computer Science and Information Technology*, 12(4):18–35.

Bharathi Raja Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Nancy Chen, Chongjia Ni, I-Fan Chen, Sunil Sivadas, Tung Pham, Haihua Xu, Xiong Xiao, Tze Lau, Su-Jun Leow, Boon Pang Lim, Cheung-Chi Leung, Chin-Hui Lee, Alvina Goh, Eng Chng, Bin Ma, and Haizhou Li. 2015. Low-resource keyword search strategies for tamil. volume 2015.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Minni Jain, Ravneet Punia, and Ishika Hooda. 2020. Neural machine translation for tamil to english. *Journal of Statistics and Management Systems*, 23(7):1251–1264.

Rachana K, Prajnashree M, Asha Hegde, and H. L Shashirekha. 2023. MUCS@DravidianLangTech2023: Sentiment analysis in code-mixed Tamil and Tulu texts using fastText. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 258–265, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

R Ramesh Kannan, Ratnavel Rajalakshmi, and Lokesh Kumar. 2021. Indicbert based approach for sentiment analysis on code-mixed tamil tweets. In *FIRE (Working Notes)*, pages 729–736.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Lal Khan, Atika Qazi, Hsien-Tsung Chang, Mousa Alhajlah, and Awais Mahmood. 2025. Empowering urdu sentiment analysis: an attention-based stacked cnn-bi-lstm dnn with multilingual bert. *Complex & Intelligent Systems*, 11(1):10.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Preprint*, arXiv:1408.5882.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Muhammad Kashif Nazir, CM Nadeem Faisal, Muhammad Asif Habib, and Haseeb Ahmad. 2025. Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages. *IEEE Access*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Suba Sri Ramesh Babu. 2022. *Sentiment Analysis In Tamil Language Using Hybrid Deep Learning Approach*. Ph.D. thesis, Dublin, National College of Ireland.

Pradeep Kumar Roy and Abhinav Kumar. 2021. Sentiment analysis on tamil code-mixed text using bi-lstm. In *FIRE (Working Notes)*, pages 1044–1050.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18.

Koyyalagunta Krishna Sampath and M Supriya. 2024. Transformer based sentiment analysis on code mixed data. *Procedia Computer Science*, 233:682–691.

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2021. Thamizhimorph: A morphological parser for the tamil language. *Machine Translation*, 35:1–34.

Model	Tokenizer	Learning Rate	Epochs	Batch Size	Max Length	Warmup Steps
mBERT	WordPiece	5e-6	17	8	256	500
XLM-R	Byte Pair Encoding (BPE)	1e-6	20	8	256	500
T5	SentencePiece	2e-4	12	8	256	500

Table A.1: Hyperparameters of transformer-based models for sentiment classification

A Tuned Hyperparameters

Table A.1 lists the hyperparameters used in transformer-based models, such as mBERT, XLM-R, and T5. Their differences include tokenization methods, learning rates, and training configurations. The batch size, maximum sequence length, and warm-up steps of all transformer models are 8, 256, and 500, respectively, to maintain stable learning. These tables together present a comparative view of the experimental setup that enables understanding of how different deep learning and transformer architectures were fine-tuned for Tamil sentiment classification.

B Class Distribution

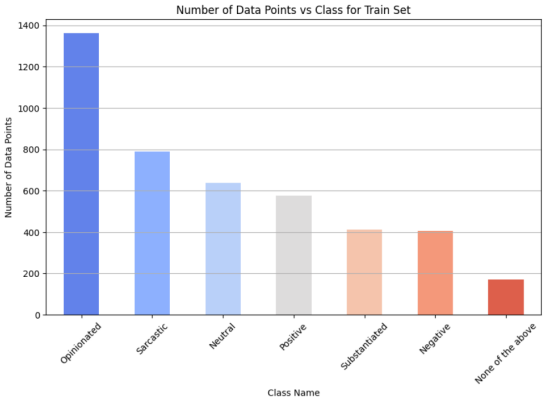


Figure B.1: Number of datapoints of each class in train dataset

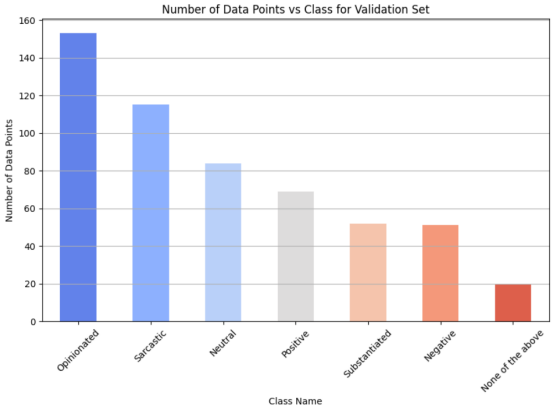


Figure B.2: Number of datapoints of each class in validation dataset

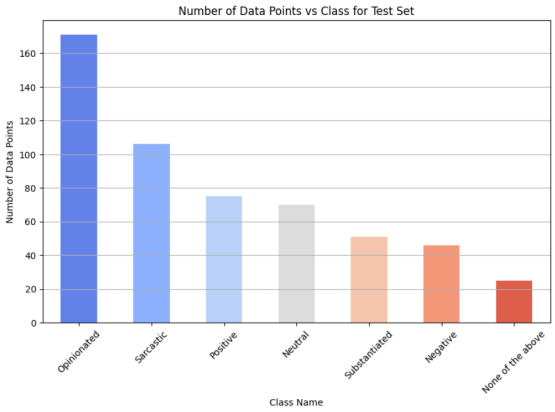


Figure B.3: Number of datapoints of each class in test dataset

The Figures B.1, B.2, and B.3 demonstrate the number of data points for each class in the training, validation, and test set, respectively.

cantnlp@DravidianLangTech 2025: A Bag-of-Sounds Approach to Multimodal Hate Speech Detection

Sidney Wong
Geospatial Research Institute
University of Canterbury
sidney.wong@pg.canterbury.ac.nz

Andrew Li
Lake Washington School District
landrewi@hotmail.com

Abstract

This paper presents the systems and results for the Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) shared task at the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025). We took a 'bag-of-sounds' approach by training our hate speech detection system on the speech (audio) data using transformed Mel spectrogram measures. While our candidate model performed poorly on the test set, our approach offered promising results during training and development for Malayalam and Tamil. With sufficient and well-balanced training data, our results show that it is feasible to use both text and speech (audio) data in the development of multimodal hate speech detection systems.

1 Introduction

There has been increased recognition within the research field that forms of hate speech on social media are not restricted to written modalities of language, but also spoken (Chhabra and Vishwakarma, 2023) and non-linguistic (i.e., memes) modalities as well (Kiela et al., 2020). As part of Multimodal Social Media Data Analysis in Dravidian Languages shared task (Lal G et al., 2025), we propose taking a 'bag-of-sounds' approach - analogous to the bag-of-words models - to train our automatic hate speech detection system on the speech (audio) data. We do this by transforming the speech (audio) data into Mel spectrogram measures and training our classification model on the outputs.

2 Related Works

The earliest automatic hate speech detection systems relied on different linguistic features such as lexical and syntactic representations (Chen et al., 2012), template-based and parts-of-speech (POS) tagging (Warner and Hirschberg, 2012), topic-modelling (Xiang et al., 2012), or a combination of lexical, POS, character bigram and term frequency-inverse document frequency (Tf-idf) representations (Dinakar et al., 2012). With a focus on hate speech in English, the model performance of these early systems yielded moderate results with limited applications to other language conditions (Jahan and Oussalah, 2023).

The introduction of transformer-based Large Language Models (LLMs), such as BERT (Devlin et al., 2019), saw an increase of word embedding feature representations jointly with neural network models in the development of hate speech detection systems (Jahan and Oussalah, 2023). Hate speech systems are now treated as a text classification task following a standardised pipeline including data set collection and labelling, feature extraction, model learning and development, and evaluation on a multiclass or binary output (Rawat et al., 2024). Both statistical language models and LLMs are used in the development of contemporary state-of-the-art hate speech detection systems.

As with other strands of Computational Linguistics and Natural Language Processing (NLP) for social impact (Hovy and Spruit, 2016), there has been a long standing tradition of shared tasks detecting hate speech and offensive language in Indo-Aryan and Dravidian languages (Chakravarthi et al., 2021; Chakravarthi et al., 2022). The best performing models in Chakravarthi et al. (2024a) were developed using open-source multilingual transformer-based LLMs (Conneau et al., 2020; Khanuja et al., 2021). In addition to testing the usability of transformed-based LLMs in non-English conditions, these systems interrogate the efficacy of text classification in code-switching (Yasaswini et al., 2021) and script-switching (Wong and Durward, 2024) phenomena.

While previous shared tasks have focused solely on written expressions of hate speech, the first multimodal social media data analysis in Dravidian languages (MSMDA-DL) was organised by

Chakravarthi et al. (2024b) with written and spoken social media language data from YouTube. The shared task provided training data with utterances in two Dravidian languages - Malayalam and Tamil - and annotated for hate speech and abusive language in YouTube videos. Only two systems were submitted as part of Chakravarthi et al. (2024b): Rahman et al. (2024) and S et al. (2024). Of interest to the current paper, Rahman et al. (2024) extracted Mel-frequency spectrogram and Mel-frequency Cepstral Coefficients (MFCCs) as acoustic features which they incorporated in their ConvLSTM.

A spectrogram is a visual representation of the acoustic frequency - or the number of vibrations in a sound wave per second - in a speech (audio) signal. The Mel-scale spectrogram, also known as Mel-frequency spectrograms or simply Mel spectrograms, is a transformation of linear machine-readable frequency measures of a spectrogram to a non-linear Mel scale which is the perceptual scale of pitch by human listeners (Stevens et al., 1937). Mel spectrograms and MFCCs (Davis and Mermelstein, 1980) have been widely used in automatic speaker recognition systems and subjective tasks such as speaker emotion recognition (Zhou et al., 2019); most recently, these acoustic measures have been included in various forms of NLP classification tasks (Arróniz and Kübler, 2023).

Rahman et al. (2024) took a Convolutional Neural Network (CNN) with Long-short term memory (LSTM), or ConvLSTM, and a hybrid 3D-CNN with LSTM approach in the development of their multimodal hate speech detection system incorporating visual, audio, and text representations. Although the shared task included training data for three Dravidian languages, only one system was designed for Tamil. During the model development phase, the system achieved a macro average F_1-score of 0.71 for Tamil. The system ranked first for Tamil in the shared task with a macro average F_1-score of 0.7143 in the test set. S et al. (2024) did not incorporate the speech (audio) components in the development of their detection system.

3 Data

The training data contained both text and speech (audio) data in three Dravidian languages: Malayalam, Tamil, and Telugu (Sreelakshmi et al., 2024). The target class labels were organised hierarchically including a binary classification with labels 'Hate' (H) and 'No Hate' (N), and a multiclass classification with five categories included Caste-related hate-speech (C), and Offensive (O), Racist (R), Sexist (S) language, and one residual non-hate speech category (N). The distribution of the target class labels in the training data for the binary classification is presented in Table 1 and for the multiclass classification in Table 2. There is significant class imbalance between target class labels between language conditions and within the training data. In addition to the class labels, the text and speech (audio) observations were identified by subject, binary gender of the speakers, source of utterance, and utterance number. We split the training data set into training and validation set, where the training set with the target labels was used to train the models and the validation set was reserved for performance evaluation.

Table 1: Binary Class Labels

Class	Malayalam	Tamil	Telugu
H	477	227	358
N	406	287	198

Table 2: Multiclass Class Labels

Class	Malayalam	Tamil	Telugu
C	186	65	122
N	406	287	198
P	118	33	58
R	91	61	72
G	82	68	106

4 Methodology

The primary purpose of the Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) shared task was to develop a hate-speech detection system that can analyse the text and speech components and predict the respective labels for three Dravidian languages: Malayalam, Tamil, and Telugu. Therefore, we approached this shared task as a classification problem. We trained a suite of candidate multimodal hate speech detection system using a statistical language model approach. While transformer-based LLMs are the state-of-the-art models in hate speech detection (Chakravarthi et al., 2024a), there is limited published research testing the use of LLMs in signal processing (i.e., audio data) as existing LLMs, such as BERT (Devlin et al., 2019), are trained on word embeddings from written language data (Verma

Table 3: Macro average F_1-score on validation training data (Binary).

	NB		SVM		LR		RF	
	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
Malayalam	0.87	0.70	0.85	0.64	0.84	**0.90**	0.84	0.91
Tamil	0.72	0.57	**0.79**	0.64	0.76	0.67	0.72	0.75
Telugu	0.64	0.60	0.66	0.65	**0.70**	0.67	0.72	0.72

Table 4: Macro average F_1-score on validation training data (Multiclass).

	NB		SVM		LR		RF	
	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
Malayalam	0.32	0.34	**0.54**	0.45	**0.54**	**0.54**	0.33	0.48
Tamil	0.14	0.21	**0.46**	0.15	0.38	0.39	0.23	0.32
Telugu	0.24	0.28	**0.55**	0.33	0.53	0.38	0.41	0.34

and Pilanci, 2024). This means we cannot directly compare the model performance of text-trained or speech-trained detection systems for the purposes of this shared task. We evaluated the performance of each candidate system according to the macro average F_1-score on the training validation data before selecting and submitting the best performing candidate model. The associated code notebook and submission data can be found at the associated GitHub repository[1].

4.1 Data Preprocessing and Feature Engineering

Multimodal hate speech data is a feature of the current shared task. Prior to the model training process, we carried out the following data preprocessing and feature engineering procedures for the two modalities:

Text: The text data was supplied in the respective Indic (Brahmic) orthographies of each language condition. We applied minimal data preprocessing on the text data as we wanted to preserve the linguistic features between the text and speech (audio) data. `CountVectorizer` and `TfidfTransformer` from `sklearn.feature_extraction.text` were employed to transform text data into numerical feature vectors suitable for machine learning models.

Speech (Audio): Mel spectrograms were computed for the audio files and converted into decibel units. To ensure uniformity across inputs, all spectrograms were padded to the same shape and reshaped into flat 2D arrays for compatibility with machine learning algorithms. Additional data normalization was implemented to meet the requirements of the Multinomial Naïve Bayes algorithm, one of the machine learning algorithms analysed in this research. This process transforms the feature matrix to ensure all feature values are scaled to the range [0,1].

4.2 Model Training, Evaluation, and Selection Criteria

The training data was split into training and validation sets with a train:test split of 75:25. Four classification models were trained for both binary and multiclass classification tasks across all three Dravidian languages. The binary classification task served as a benchmark. The statistical methods used were as follows: Multinomial Naïve Bayes (NB), Linear Support Vector Machine (SVM), Logistic Regression Classifier (LR), and Random Forest Classifier (RF). There were 48 candidate models in total according to the following rubric: two classifications X three language conditions X two modalities X four statistical methods.

The performance of each candidate model was evaluated by macro average F_1-score. The model performance as measured by macro average F_1-score for the binary classification models are presented in Table 3 and for the multiclass classification models in Table 4. The best performing model for each language condition (row) is highlighted in **bold**. For completeness and benchmarking purposes, we have also included the model performance based on F_1-score for binary and multiclass classification models in the Appendix as shown in Tables 7 to 12.

For model evaluation, we used a macro average

[1] https://github.com/sidneygjwong/cantnlp-dravidianlangtech2025

Table 5: macro average F_1-score from test evaluation and rank by language.

Language	F_1-score	Rank
Malayalam	0.273	14
Tamil	0.3186	9
Telugu	0.1774	12

F_1 score as the primary metric. Overall, Logistic Regression (LR) achieved the highest macro average F_1-score and performed the best among all four algorithms for speech (audio) data in the binary classification candidate models as shown in Table 3. In contrast, Linear SVM (SVM) had better performance with the multiclass classification candidate models as shown in Table 4. We notice a significant drop in performance between the binary classification to the multiclass classification models with the maximum macro average F_1-score of 0.90 lowering to 0.54.

Even though the text data trained Linear SVM (SVM) models largely outperformed the speech (audio) trained candidate models, we opted for the best performing multiclass speech (audio) data trained models. We justify this decision as the purpose of the shared task was to incorporate multimodal language data and not just one modality. Where text data only encodes linguistic information, we argue speech (audio) data implicitly encodes both linguistic and paralinguistic features of hate speech. Furthermore, the performance of the text trained models Linear SVM (SVM) models only performed marginally better than our optimal model - the speech (audio) trained logistic regression (LR) models.

5 Results

The macro average F_1-scores of our candidate model on the test set are presented in Table 5. All three models performed poorly on the test set with a macro average F_1-score below chance. In contrast, the best performing model in Malayalam and Tamil was by Team `SSNTrio` who yielded a macro average F_1-score of 0.7511 for Malayalam and 0.7332 for Tamil. The best performing model in Telugu was by Team `lowes` had a macro average F_1-score of 0.3817.

6 Discussion

Based on the evaluation metrics alone, our optimal method performed poorly across all three language conditions with a macro average F_1-score below chance. With reference to Table 4 and Table 5, we see a significant drop in performance between the validation evaluation and test evaluation. This drop is particularly stark in Malayalam where the macro average F_1-score went from 0.54 to 0.273, and for Telugu from 0.38 to 0.1774. Malayalam had a median macro average F_1-score of 0.41 and average of 0.38; for Tamil a median of 0.32 and average of 0.34; and for Telugu a median of 0.24 and an average of 0.23.

Table 6: Relative proportion of test ($n = 10$) to train data as a percentage (%) per class label.

Class	Malayalam	Tamil	Telugu
C	5.4	15.4	8.2
N	2.5	3.5	5.1
P	8.5	30.3	17.2
R	11.0	16.4	13.9
G	12.2	14.7	9.4

While the median and mean scores suggest an improvement from Chakravarthi et al. (2024b) across the board, we argue the consistently poor performance may indicate there are underlying issues with the training data. One possible explanation for the poor model performance is the class imbalance observed in Table 2 where we see not only difference in utterances between language conditions, but also between classes especially in the minority classes. When we consider the relative proportion of utterances in the test evaluation set as shown in Table 6, some utterances in the minority classes are over represented while utterances in the majority classes are under represented.

As we refer back to the binary classification models as shown in Table 3 (which were not part of the shared task), we can see that the models performed well not only across all language conditions, but also across the different statistical methods. Even though the models performed poorly on the test set which suggests some modifications are needed to our pipeline, some of the poor model performance can be attributed to the class imbalance in the training and test data. It is possible the decline in performance from the validation to test data suggests possible over-fitting to the training set or other dataset related biases not accounted for in the current model development pipeline which will be worthy of further investigation.

Possible improvements to our existing model

may include further hyper-parameter tuning such as employing optimisation techniques such as GridSearchCV or RandomisedSearchCV to fine-tune parameters for the text classifiers used in our study such as Random Forest, SVM, and Logisitic Regression. We could explore more advanced boosting algorithms like XGBoost, CatBoost, or LightGBM which may improve classification performance. Alternatively, we could look into comparing other speech feature representations such as Mel-Frequency Cepstral Coefficients (MFCCs) in addition to Mel spectrograms which have also been effective in speech-based classification tasks.

The current study provides a foundation for future work in the development of multimodal hate speech detection systems. Despite the lower than expected performance of our proposed approach when compared to other teams in the shared task, we demonstrated in this paper that speech data carries valuable extra-linguistic information for hate speech detection. We argue that further improvements in training data representation and model architecture (i.e., with state-of-the-art methodologies) may yield better performance.

7 Conclusion

While our candidate model performed poorly on the test set, our 'bag-of-sounds' approach offered promising results during training and development for Malayalam and Tamil. With sufficient and well-balanced training data, our results show that it is feasible to use both text and speech (audio) data in the development of multimodal hate speech detection systems. It is important to note that our current study intentionally avoided state-of-the-art deep learning or large language models to avoid overloading our existing approach with speculative enhancements from deep learning and transformer-based language models; however, we will look to incorporate more sophisticated models, such as ELMo (Peters et al., 2018), in future studies to determine the performance of our proposed approach alongside state-of-the-art methodologies.

Acknowledgements

We would like to thank the four anonymous peer reviewers and the organisers of the Fifth Workshop on Speech and Language Technologies for Dravidian Languages (DravidianLangTech-2025) co-located at the North American Chapter of the Association for Computational Linguistics in Albuquerque, New Mexico. We would also like to acknowledge Fulbright New Zealand | Te Tūāpapa Mātauranga o Aotearoa me Amerika and their partnership with the Ministry of Business, Innovation, and Employment | Hīkina Whakatutuki for their support through the Fulbright New Zealand Science and Innovation Graduate Award.

References

Santiago Arróniz and Sandra Kübler. 2023. Was That a Question? Automatic Classification of Discourse Meaning in Spanish. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 132–142, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Rajat Kumar Behera, Pradip Kumar Bala, Nripendra P. Rana, and Zahir Irani. 2023. Responsible natural language processing: A principlist framework for social benefits. *Technological Forecasting and Social Change*, 188:122306.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024a. Overview of Third Shared Task on Homophobia and Transphobia Detection in Social Media Comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of The Shared Task on Homophobia and Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the Shared Task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Sajeetha Thavareesan, Elizabeth Sherly, Rajeswari Nadarajan, and Manikandan Ravikiran. 2024b. Findings of the Shared Task on Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL)@DravidianLangTech

2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61, St. Julian's, Malta. Association for Computational Linguistics.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

S. Davis and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366. Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Trans. Interact. Intell. Syst.*, 2(3):18:1–18:30.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages. *arXiv preprint*. ArXiv:2103.10730 [cs].

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Salla-Maaria Laaksonen, Jesse Haapoja, Teemu Kinnunen, Matti Nelimarkka, and Reeta Pöyhtäri. 2020. The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring. *Frontiers in Big Data*, 3.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Albuquerque, NM. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate Speech Classifiers are Culturally Insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical Considerations in NLP Shared Tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Md. Rahman, Abu Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das, and Mohammed Moshiul Hoque. 2024. Binary_beasts@DravidianLangTech-EACL 2024: Multimodal Abusive Language Detection in Tamil based on Integrated Approach of Machine Learning and Deep Learning Techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and*

Language Technologies for Dravidian Languages, pages 212–217, St. Julian's, Malta. Association for Computational Linguistics.

Anchal Rawat, Santosh Kumar, and Surender Singh Samant. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Computational Statistics*, 16(2):e1648.

Anierudh S, Abhishek R, Ashwin Sundar, Amrit Krishnan, and Bharathi B. 2024. Wit Hub@DravidianLangTech-2024:Multimodal Social Media Data Analysis in Dravidian Languages using Machine Learning Models. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 229–233, St. Julian's, Malta. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of Hate Speech and Offensive Language CodeMix Text in Dravidian Languages Using Cost-Sensitive Learning Approach. *IEEE Access*, 12:20064–20090. Conference Name: IEEE Access.

S. S. Stevens, J. Volkmann, and E. B. Newman. 1937. A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

Prateek Verma and Mert Pilanci. 2024. Towards Signal Processing In Large Language Models. *arXiv preprint*. ArXiv:2406.10254 [cs] version: 1.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, USA. Association for Computational Linguistics.

Sidney Wong. 2024. Sociocultural Considerations in Monitoring Anti-LGBTQ+ Content on Social Media. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 84–97, Bangkok, Thailand. Association for Computational Linguistics.

Sidney Wong and Matthew Durward. 2024. cantnlp@LT-EDI-2024: Automatic Detection of Anti-LGBTQ+ Hate Speech in Under-resourced Languages. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 177–183, St. Julian's, Malta. Association for Computational Linguistics.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 1980–1984, New York, NY, USA. Association for Computing Machinery.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer Learning for Offensive Language Detection in Dravidian Languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194, Kyiv. Association for Computational Linguistics.

Hengshun Zhou, Debin Meng, Yuanyuan Zhang, Xiaojiang Peng, Jun Du, Kai Wang, and Yu Qiao. 2019. Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition. In *2019 International Conference on Multimodal Interaction*, ICMI '19, pages 562–566, New York, NY, USA. Association for Computing Machinery.

A Limitations

While we saw promising results in the 'bag-of-sounds' approach we proposed in the development of our hate speech detection system; there are two main limitations we wish to address in our system. The first being the use of statistical language models; and secondly, the lack of sociolinguistic input in the development of our model.

Firstly, our candidate does not use state-of-the-art modelling in the development of our system as we have not incorporated transformer-based LLMs. While we justified our reasons for excluding LLMs in Section 4 in order to maintain comparability between the two modalities, we need to determine how we might incorporate the use of LLMs in our system as existing state-of-the-art models rely on multilingual LLMs (Chakravarthi et al., 2024a). With the development of LLMs for speech (audio) signal processing (Verma and Pilanci, 2024), it may be possible for us to replicate our analysis. As we will discuss in the Ethics Statement, introducing LLMs may inadvertently introduce bias into our hate speech detection system.

This leads us to discuss the second limitation which is the lack of sociolinguistic input in the development of our system. As discussed in Section 3, each utterance was labelled for one demographic variable - the binary gender classification of the speaker. While we are aware that speech varies based on gender as a result of mechanical and social demographic differences, we have not incorporated in the development of our model. This means

Table 7: Model comparison metrics by F_1-score per class (binary) in Malayalam

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
H	0.89	0.71	0.87	0.78	0.86	0.92	0.87	0.92
N	0.86	0.69	0.84	0.49	0.82	0.89	0.82	0.90

Table 8: Model comparison metrics by F_1-score per class (multiclass) in Malayalam

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
C	0.72	0.44	0.80	0.73	0.68	0.70	0.70	0.63
N	0.00	0.20	0.21	0.08	0.35	0.31	0.00	0.15
P	0.73	0.71	0.85	0.89	0.81	0.90	0.73	0.84
R	0.14	0.19	0.38	0.55	0.42	0.58	0.13	0.45
G	0.00	0.16	0.44	0.00	0.46	0.21	0.08	0.35

Table 9: Model comparison metrics by F_1-score per class (binary) in Tamil

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
H	0.64	0.49	0.77	0.63	0.73	0.61	0.66	0.70
N	0.80	0.64	0.81	0.65	0.78	0.73	0.77	0.79

Table 10: Model comparison metrics by F_1-score per class (multiclass) in Tamil

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
C	0.00	0.00	0.16	0.00	0.07	0.08	0.11	0.10
N	0.00	0.15	0.47	0.00	0.49	0.35	0.32	0.21
P	0.70	0.69	0.81	0.74	0.76	0.72	0.72	0.79
R	0.00	0.20	0.33	0.00	0.17	0.43	0.00	0.22
G	0.00	0.00	0.53	0.00	0.39	0.36	0.00	0.26

Table 11: Model comparison metrics by F_1-score per class (binary) in Telugu

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
H	0.84	0.74	0.78	0.73	0.80	0.81	0.85	0.81
N	0.44	0.46	0.54	0.57	0.60	0.52	0.58	0.52

Table 12: Model comparison metrics by F_1-score per class (multiclass) in Telugu

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
C	0.34	0.44	0.65	0.56	0.57	0.55	0.65	0.50
N	0.33	0.14	0.59	0.28	0.52	0.34	0.47	0.27
P	0.51	0.43	0.65	0.53	0.67	0.49	0.68	0.57
R	0.00	0.32	0.36	0.13	0.38	0.24	0.12	0.37
G	0.00	0.07	0.52	0.17	0.48	0.26	0.12	0.00

it is possible that there are unexplained predictors in the data that are unaccounted for namely acoustic differences between these social categories such as gender, culture, and possibly geographic dialect bias (Wong, 2024). Therefore, future work should involve more in-depth analysis on the speech (audio) data which may justify the need to further normalise or standardise the data.

B Ethics/Broader Impact Statement

Parra Escartín et al. (2017) argued that shared tasks play an important role in Computational Linguistics and Natural Language Processing (NLP) as it helps encourage a culture within the field to develop upon the state-of-the-art. With an increased recognition of automatic hate speech detection beyond just written text to other modalities of language (Chhabra and Vishwakarma, 2023), the current shared task plays an important role in how detection can be achieved in not only multimodal but also multilingual contexts.

In spite of these benefits, there are also ethical issues and negative effects of competition in NLP shared tasks such as secretive behaviour, overlooking the relevance of ethical concerns, unconscious overlooking of ethical concerns, redundancy and replicability in the field among other concerns (Parra Escartín et al., 2017). With the 'datafication' of hate speech an increasing issue within the field of hate speech detection (Laaksonen et al., 2020), we will consider the ethics and broader impacts of our proposed system within the context of the current shared task guided by the eight principals of *Responsible NLP* (Behera et al., 2023).

Principal 1: Well-being The current system contributes to our current understanding of multimodal automatic hate speech detection in three Dravidian languages. The current system was designed alongside junior researchers which supports development in working with low- and under-resourced language condition often overlooked in NLP research.

Principal 2: Human-Centred Values The current system does not include human subjects, external annotators, or additional data from external sources. However, this is also an area of improvement where the researchers can work alongside target communities - namely Malayalam, Tamil, and Telugu speakers - in the development of a system that is fit for purpose.

Principal 3: Fairness While we have avoided to the best of our ability to not perpetuate existing prejudice towards marginalised and vulnerable communities, we are aware that hate speech training datasets are sensitive to racial (Davidson et al., 2019; Sap et al., 2019) and sociocultural (Lee et al., 2023; Wong, 2024). Therefore, we propose that further work is needed to determine the presence of underlying biases within the training data and possible downstream impacts of these biases.

Principal 4: Privacy and Security In accordance with the terms and conditions of the shared task, the authors have not re-distributed the data and have only used the data for non-commercial and academic-research purposes. We have not used the data for surveillance, analyses, or research that isolates a group of individuals for unlawful or discriminatory purposes.

Principal 5: Reliability We have provided the model performance metrics which can be found throughout the paper and in the Appendix. We acknowledge there will be variances within the metrics due to the stochastic nature of statistical language models. There is limited risk to organisers, authors, or users who wish to reproduce our systems.

Principal 6: Transparency We have described our system to the best of our ability for other researchers to reproduce our system; however, we are limited to the metadata provided of the training data provided to us by the organisers of the shared task. We have not involved additional human subjects or external annotators.

Principal 7: Interrogation We encourage readers to refer to the other system description papers associated with this shared task.

Principal 8: Accountability We encourage readers to contact the authors to discuss the contents of this paper.

LexiLogic@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages

Billodal Roy[1], Pranav Gupta[1], Souvik Bhattacharyya[1], Niranjan Kumar M[1]
[1]Lowe's
Correspondence: {billodal.roy, pranav.gupta, souvik.bhattacharyya, niranjan.k.m}@lowes.com

Abstract

This paper describes our participation in the DravidianLangTech@NAACL 2025 shared task on hate speech detection in Dravidian languages. While the task provided both text transcripts and audio data, we demonstrate that competitive results can be achieved using text features alone. We employed fine-tuned Bidirectional Encoder Representations from Transformers (BERT) models from l3cube-pune for Malayalam, Tamil, and Telugu languages. Our system achieved notable results, securing first position for Telugu, and second position for Tamil and Malayalam tasks in the official leaderboard.

1 Introduction

The increasing volume of social media content in Dravidian languages has heightened the need for robust hate speech detection systems. The DravidianLangTech shared task at NAACL 2025 presented a multimodal challenge for hate speech detection in Malayalam, Tamil, and Telugu (Lal G et al., 2025; Premjith et al., 2024a,b; Sreelakshmi et al., 2024). These languages, with their rich morphological structures and distinct scripts, present unique challenges for automated content moderation. Our work demonstrates that while multimodal approaches are valuable, significant performance can be achieved through focused analysis of textual content alone. We utilized language-specific BERT (Devlin et al., 2019) models from l3cube-pune (Joshi, 2023), fine-tuned on the text transcripts from the task-specific dataset. This approach not only proved computationally efficient but also highly effective, suggesting that textual features capture substantial indicators of hate speech in these languages.[1]

[1]The code for this work is available at https://github.com/prannerta100/naacl2025-dravidianlangtech

2 Related Work

Research in hate speech and offensive language detection has evolved significantly, particularly for social media content. Initial approaches relied on traditional machine learning methods, utilizing handcrafted features such as n-grams and sentiment lexicons (Sreelakshmi et al., 2020). These methods, while foundational, faced limitations in capturing the contextual complexities of natural language, especially in code-mixed and low-resource scenarios.

The emergence of transformer architectures marked a significant advancement in this domain. The introduction of BERT (Devlin et al., 2019) enabled more sophisticated contextual representations, leading to substantial improvements in detection accuracy. Building on this foundation, Chakravarthi et al. (Chakravarthi et al., 2023) demonstrated enhanced performance by combining MPNet with convolutional neural networks for Dravidian languages, specifically addressing code-mixing challenges in Tamil, Malayalam, and Kannada. This work was complemented by Subramanian et al. (Subramanian et al., 2022), who focused on Tamil YouTube comments, highlighting the importance of handling class imbalance in social media content. Multilingual approaches have further advanced the field through innovative architectures. Hande et al. (Hande et al., 2022) explored multi-task learning with mBERT, simultaneously addressing sentiment analysis and offensive language detection. Roy et al. (Roy et al., 2022) proposed an ensemble framework that integrates multiple approaches, demonstrating the advantages of combining traditional and modern methodologies.

Recent work has increasingly focused on language-specific adaptations. Notable contributions include Pillai and Arun's (Pillai and Arun, 2024) investigation of feature fusion techniques for Malayalam, and IIITDWD-ShankarB's (Biradar

Category	Malayalam	Tamil	Telugu
Non-Hate	406	287	198
Gender	82	68	106
Political	118	33	58
Religious	91	61	72
Character	186	65	122
Total	883	514	556

Table 1: Distribution of instances across categories for each language in the dataset

and Saumya, 2022) application of mBERT to South Indian languages. Arunachalam et al. (Arunachalam and Maheswari, 2024) demonstrated the effectiveness of language-specific BERT models in achieving competitive performance using only textual features. Additional research by Sharma et al. (Sharma et al., 2023) on detecting specific forms of discriminatory content has further emphasized the importance of language-tailored approaches. This progression in the field reflects a clear shift from feature-engineered solutions to sophisticated transformer-based systems, better equipped to handle the nuances of code-mixed content and class imbalance in Dravidian language hate speech detection.

3 Dataset and Task Description

The DravidianLangTech shared task provided datasets for hate speech detection in three Dravidian languages: Malayalam, Tamil, and Telugu. Each dataset consists of text transcripts and audio recordings sourced from YouTube videos, categorized into hate and non-hate speech, with hate speech further subdivided into four categories.

3.1 Data Organization

The data follows a structured format with detailed file nomenclature containing speaker information, source identifiers, and classification labels. Each instance includes both audio recording and corresponding text transcript, though our approach utilizes only the text components.

3.2 Class Distribution

Table 1 shows the distribution of instances across different categories for each language.

3.3 Data Characteristics

A notable characteristic of the dataset is its class imbalance, with Non-Hate being the dominant category across all three languages. The distribution of hate speech subcategories varies significantly among languages, with Character Defamation being particularly prevalent in Malayalam and Telugu datasets. This imbalanced distribution presents a significant challenge for model training and necessitates careful consideration during the development of our classification approach.

4 Methodology

Our approach leverages language-specific BERT models fine-tuned for each Dravidian language, with a focus on optimizing for the inherent class imbalance in the dataset.

4.1 Model Architecture

We utilized pre-trained BERT models from l3cube-pune, specifically tailored for Dravidian languages. These models have demonstrated superior performance in capturing language-specific nuances compared to general multilingual models. The base architecture consists of the pre-trained BERT model with a classification head fine-tuned for our specific task.

Language	Base Model
Malayalam	l3cube-pune/malayalam-bert
Tamil	l3cube-pune/tamil-bert
Telugu	l3cube-pune/telugu-bert

Table 2: Language-specific BERT models

4.2 Implementation Details

The implementation utilized the Hugging Face Transformers library for model architecture and training. We maintained the original text without pre-processing, allowing the models to learn from the natural language patterns. The system was implemented using PyTorch, with training facilitated by the Transformers library's Trainer API.

Parameter	Value
Learning Rate	2e-5
Batch Size	8
Training Epochs	15-20
Label Smoothing	0.1
Weight Decay	0.005-0.01

Table 3: Training hyperparameters

4.3 Training Strategy

Our training approach evolved through systematic experimentation. Initially, we employed an 80-20 train-test split while maintaining class distribution. To address the class imbalance, we implemented label smoothing and weight decay regularization. The final models were trained on the complete dataset after parameter optimization, achieving robust performance across all categories. The training process incorporated early stopping based on evaluation loss to prevent overfitting, along with model checkpointing to retain the best-performing version. We found that a learning rate of 2e-5 with a batch size of 8 provided optimal convergence across all three languages, though Telugu required slightly higher weight decay for better generalization.

5 Results and Analysis

Our system demonstrated competitive performance across all three languages in the DravidianLangTech shared task. We achieved second rank in Malayalam and Tamil tasks, and first rank in Telugu, showcasing the effectiveness of our approach.

Language	Macro F1	Rank
Malayalam	0.7367	2/17
Tamil	0.7225	2/17
Telugu	0.3817	1/18

Table 4: Final test set performance and rankings

In the Malayalam task, our system achieved a macro F1 score of 0.7367, placing second behind SSNTrio (0.7511). The margin between the top two systems was relatively small (0.0144), indicating comparable performance levels. For Tamil, we again secured the second position with a macro F1 score of 0.7225, closely following SSNTrio (0.7332). In the Telugu task, our system outperformed all other participants with a macro F1 score of 0.3817, marginally ahead of SSNTrio (0.3758).

5.1 Error Analysis and Model Behavior

Our development set experiments showed notably different performance patterns compared to the final test set results, highlighting important insights about model generalization. During development, the Malayalam model achieved a macro F1 score of 0.80 on our test split, significantly higher than the 0.7367 obtained on the competition's test set. This performance gap suggests potential over-fitting despite our regularization efforts.

The most pronounced discrepancy appeared in the Telugu task. While our development experiments showed exceptional performance with a macro F1 score of 0.90, the final test set yielded 0.3817. This substantial difference indicates that the competition's test data likely contained more challenging or diverse examples than our training split. However, it's noteworthy that this performance level still led to a first-place ranking, suggesting that other teams faced similar generalization challenges.

The Tamil model showed the most consistent performance between development (0.52 macro F1) and final test set (0.7225) results. This consistency might be attributed to our more conservative hyperparameter choices for Tamil, particularly in terms of regularization strength.

Across all languages, we observed that the models performed most reliably on non-hate speech classification, likely due to the larger representation of this class in the training data. The detection of political hate speech proved particularly challenging, especially in Tamil where the training data was most limited for this category. These observations suggest that while our approach effectively captures general language patterns, performance on minority classes remains sensitive to data distribution shifts between training and test sets.

6 Discussion and Future Directions

The significant performance variations between our development experiments and the final test set results highlight key areas for improvement in our approach. While achieving competitive rankings, the disparity (particularly in Telugu with 0.90 in development vs 0.3817 in final test) indicates a need for more robust validation strategies.

To enhance model performance, we recommend implementing language-specific data augmentation techniques and adopting more rigorous cross-validation approaches. Our text-only implementation, while competitive, could benefit from integrating the available audio features. Recent advances in speech encoders for Indian languages, such as IndicWav2Vec (Javed et al., 2021), could provide valuable additional signals for hate speech detection.

Furthermore, focusing on Dravidian language-specific characteristics through better morphologi-

cal analysis and script handling could improve the model's understanding of regional language variations. As larger language models trained specifically on Dravidian languages become available, they may offer better feature representations for this task. These improvements, combined with effective multimodal fusion strategies, could lead to more robust and generalizable models for hate speech detection in Dravidian languages.

7 Conclusion

This paper presented our approach to hate speech detection in Dravidian languages as part of the DravidianLangTech shared task at NAACL 2025. By leveraging language-specific BERT models and implementing careful optimization strategies, we achieved competitive results across all three languages, securing first position in Telugu and second positions in both Malayalam and Tamil tasks.

Our results demonstrate that transformer-based models, even without multimodal features, can effectively detect hate speech in Dravidian languages. The performance variations between development and final test sets provided valuable insights into the challenges of model generalization in this domain. The success of our text-only approach, while encouraging, suggests potential for further improvements through multimodal integration and language-specific optimizations.

8 Limitations

While this work demonstrates effective hate speech detection in Dravidian languages using language-specific BERT models, several important limitations should be acknowledged. One key limitation is the exclusive reliance on textual features, which means that the audio components available in the dataset are not utilized. Although our approach achieved competitive results, it may miss important paralinguistic cues—such as tone, emphasis, and emotion—that could provide additional context when the text alone is ambiguous.

Another limitation is related to the computational resources required for training and inference. The language-specific BERT models, while powerful, demand significant processing power, which may restrict their use in real-time content moderation scenarios where rapid processing of large volumes of data is essential.

A further challenge lies in handling class imbalance in the training data. Despite applying regularization techniques such as label smoothing and weight decay, the models still tend to favor majority classes. This bias is particularly evident in the detection of certain types of hate speech, such as political hate speech in Tamil, where training examples are limited. This suggests that the current approach might not fully capture the nuances of minority hate speech categories.

Additionally, our models face difficulties with code-mixed content—a common characteristic of social media communication in Dravidian languages. Although language-specific BERT models capture many linguistic nuances, they may not optimally process text that switches between English and the target language or different scripts, which is increasingly prevalent online.

Finally, the observed performance disparity between the development and final test sets, especially in the Telugu task, indicates limitations in the model's ability to generalize to new, unseen data. This gap suggests that the current approach may not be robust enough to handle shifts in data distribution or novel patterns of hate speech that emerge over time.

These limitations offer clear directions for future work, including the integration of multimodal features, improved techniques for handling class imbalance and code-mixed text, and the development of more robust validation and adaptation strategies.

References

V. Arunachalam and N. Maheswari. 2024. Enhanced detection of hate speech in dravidian languages in social media using ensemble transformers. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19(Article 39).

Shankar Biradar and Sunil Saumya. 2022. Iiitdwd-shankarb@dravidian-codemixi-hasoc2021: mbert based model for identification of offensive content in south indian languages. *Preprint*, arXiv:2204.10195.

Bharathi Raja Chakravarthi et al. 2023. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

A. Hande, S. U. Hegde, and B. R. Chakravarthi. 2022. Multi-task learning in under-resourced dravidian languages. *Journal of Data, Information and Management*, 4:137–165.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2021. Towards building asr systems for the next billion users. *Preprint*, arXiv:2111.03945.

Raviraj Joshi. 2023. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *Preprint*, arXiv:2211.11418.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Aditya R Pillai and Biri Arun. 2024. A feature fusion and detection approach using deep learning for sentimental analysis and offensive text detection from code-mix malayalam language. *Biomedical Signal Processing and Control*, 89:105763.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

P. K. Roy, S. Bhawal, and C. N. Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

Deepawali Sharma, Vedika Gupta, and Vivek Kumar Singh. 2023. Detection of homophobia & transphobia in dravidian languages: Exploring deep learning methods. *Preprint*, arXiv:2304.01241.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

K. Sreelakshmi, B. Premjith, and K. P. Soman. 2020. Detection of hate speech text in hindi-english code-mixed data. In *Procedia Computer Science*, volume 171, pages 737–744.

M. Subramanian, G. J. Adhithiya, S. Gowthamkrishnan, and R. Deepti. 2022. Detecting offensive tamil texts using machine learning and multilingual transformer models. In *Proceedings of the International Conference on Smart Technologies and Systems for Next Generation Computing*, pages 1–6.

LexiLogic@DravidianLangTech 2025: Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments and Sentiment Analysis in Tamil and Tulu

Billodal Roy*, **Souvik Bhattacharyya***, **Pranav Gupta, Niranjan Kumar M**
Lowe's
Correspondence: {billodal.roy, souvik.bhattacharyya, pranav.gupta, niranjan.k.m}@lowes.com

Abstract

We present our approach and findings for two sentiment analysis shared tasks as part of DravidianLangTech@NAACL 2025. The first task involved a seven-class political sentiment classification for Tamil tweets, while the second addressed code-mixed sentiment analysis in Tamil-English and Tulu-English social media texts. We employed language-specific BERT models fine-tuned on the respective tasks, specifically utilizing the L3Cube-Tamil-BERT for Tamil classification and a Telugu-based BERT model for Tulu classification. Our system achieved notable results, particularly securing the first position in the Tulu code-mixed sentiment analysis track. The experiments demonstrate the effectiveness of language-specific pre-trained models for Dravidian language sentiment analysis, while also highlighting the challenges in handling political discourse and code-mixed content.

1 Introduction

Sentiment analysis in low-resource languages presents unique challenges, particularly for Dravidian languages with their rich morphological structure and increasing prevalence of code-mixing in social media contexts. This paper presents our unified approach to two distinct but complementary sentiment analysis tasks in the DravidianLangTech@NAACL 2025 shared task.

The first task addresses the complex challenge of political sentiment analysis in Tamil tweets, requiring fine-grained classification into seven distinct categories: Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral, and None of the above. This multi-class approach enables a more nuanced understanding of political discourse compared to traditional positive-negative sentiment classifications.

The second task focuses on sentiment analysis in code-mixed scenarios(Chakravarthi et al., 2020; Hegde et al., 2022, 2023; S. K. et al., 2024), specifically Tamil-English and Tulu-English social media comments(Durairaj et al., 2025). Code-mixing, a common phenomenon in multilingual communities, introduces additional complexity due to the interplay of linguistic features from multiple languages within the same text.

Our approach leverages recent advances in transformer-based models, specifically utilizing language-specific BERT (Devlin et al., 2019) models fine-tuned for the respective tasks. For Tamil classification, we employed L3Cube-Tamil-BERT (Joshi, 2022), while for Tulu, we innovatively adapted a Telugu-based BERT model, demonstrating the potential for cross-lingual transfer in closely related languages.[1]

2 Related Works

Sentiment analysis of textual data has been a prominent area of research for many years, with product and movie reviews being among the most extensively studied topics (Wankhade et al., 2022; S. K. et al., 2024). Many platforms include a rating system alongside text input, simplifying data preparation and framing the task as a supervised learning problem. In contrast, detecting political sentiment is considerably more challenging, as it involves distinguishing not only between political and non-political content but also identifying nuances such as sarcasm or references to specific individuals or real-time events.

Political sentiment analysis has been studied using both rule-based methods and machine learning algorithms. For instance, Elghazaly et al., 2016 utilized TF-IDF to extract document vectors and applied Support Vector Machines (SVM) and Naive

*These authors contributed equally to this work.

[1]The code for this work is available at https://github.com/prannerta100/naacl2025-dravidianlangtech

Bayes classifiers to analyze the sentiment of Arabic texts. Similarly, Bakliwal et al., 2013 demonstrated that combining a simple lexicon-based approach with bag-of-words features significantly improves accuracy. Beyond traditional supervised models like SVM and Logistic Regression, Ansari et al., 2020 employed LSTM-based (Hochreiter, 1997) models using TF-IDF features of unigrams, bigrams, and trigrams.

The rise of deep learning has transformed sentiment analysis by introducing techniques like recurrent neural networks (RNNs) and, more recently, transformers (Vaswani et al., 2017), which excel at capturing context and relationships between words. Furthermore, the development of highly parameterized large language models (LLMs) (Radford et al., 2018; Touvron et al., 2023) has made it more feasible to fine-tune models for entirely new tasks, eliminating the need to train them from scratch.

3 Dataset and Task Description

This section details the datasets and specific requirements for both the political sentiment and code-mixed sentiment analysis tasks.

3.1 Dataset Statistics

The political sentiment analysis task utilized Tamil Twitter data (Chakravarthi et al., 2025), while the code-mixed task covered Tamil-English and Tulu-English social media comments (Chakravarthi et al., 2020; Hegde et al., 2022, 2023). Tables 1, 2, and 3 present the class distributions for each dataset.

Category	Train	Test	Total
Opinionated	1,361	153	1,514
Sarcastic	790	115	905
Neutral	637	84	721
Positive	575	69	644
Substantiated	412	52	464
Negative	406	51	457
None	171	20	191
Total	4,352	544	4,896

Table 1: Political Sentiment Dataset Distribution

3.2 Task Requirements

Both tasks required sentiment classification at the message level, though with distinct objectives. The political sentiment task demanded fine-grained classification into seven categories, capturing the nuanced nature of political discourse. The classification schema included substantiated opinions, sarcasm detection, and general sentiment polarity.

Category	Train	Validation	Total
Not Tulu	4,400	543	4,943
Positive	3,769	470	4,239
Neutral	3,175	368	3,543
Mixed	1,114	143	1,257
Negative	843	118	961
Total	13,301	1,642	14,943

Table 2: Tulu-English Code-Mixed Dataset Distribution

Category	Train	Validation	Total
Positive	18,145	2,272	20,417
Unknown_state	5,164	619	5,783
Negative	4,151	480	4,631
Mixed_feelings	3,662	472	4,134
Total	31,122	3,843	34,965

Table 3: Tamil-English Code-Mixed Dataset Distribution

The code-mixed task focused on handling the complexity of bilingual text while performing sentiment classification. This task presented additional challenges due to the informal nature of social media language and the intricate patterns of language mixing. For Tamil-English, systems needed to classify texts into four categories, while Tulu-English required classification into five categories.

4 Methodology

Our approach utilized transformer-based models across all tasks, specifically leveraging language-specific BERT variants. We employed a consistent fine-tuning strategy while adapting the hyperparameters and training configurations to each task's unique requirements. Additionally, we used the text data in its raw form without any preprocessing, such as handling emojis, removing stopwords, or performing normalization.

4.1 Model Architecture

For the political sentiment analysis task, we experimented with multiple transformer based encoder and decoder models: the multilingual BERT model, a monolingual Tamil BERT model developed by L3Cube, and GPT-2 (Radford et al., 2019). The models were initialized with pre-trained weights and augmented with a classification head with a 10% dropout rate.

For the code-mixed sentiment analysis tasks, we utilized language-specific BERT models. The Tamil-English classification employed the L3Cube Tamil-BERT model, while the Tulu-English classification innovatively used the L3Cube Telugu-BERT model, leveraging the linguistic similarities between Tulu and Telugu. Each model was configured with task-specific classification heads matching their respective output dimensions: four classes for Tamil-English and five classes for Tulu-English.

4.2 Training Configuration

We implemented distinct training configurations for each task to address their specific challenges:

4.2.1 Political Sentiment Analysis

The political sentiment classifier was trained using the following configuration, as detailed in Table 4:

Parameter	Value
Learning rate	5e-5
Learning rate decay	0.9
Batch size	64
Training epochs	10
Dropout rate	10%

Table 4: Political Sentiment Training Parameters

4.2.2 Code-Mixed Sentiment Analysis

For the code-mixed tasks, we implemented separate configurations for Tamil-English and Tulu-English classification, as shown in Table 5:

Param	Tamil-En	Tulu-En
Learning rate	2e-7	2e-5
Batch Size	32	16
Epochs	5	3
Weight Decay	0.005	0.005
Label Smoothing	0.1	0.1
Grad. Acc. Steps	4	1

Table 5: Code-Mixed Training Parameters

4.3 Optimization Strategies

To address the class imbalance present in both tasks, we implemented several optimization techniques. For the code-mixed tasks, we employed label smoothing with a factor of 0.1 and weight decay of 0.005. The Tamil-English model additionally utilized gradient accumulation with 4 steps to effectively increase the batch size while managing memory constraints.

For all tasks, we utilized the AdamW optimizer and implemented mixed-precision training (FP16) to improve computational efficiency. The models were trained with early stopping based on validation loss, with checkpoints saved at each epoch. To enhance training efficiency, we employed data loading optimizations including pinned memory and multi-worker data loading for the Tamil-English task.

5 Results and Discussion

We present the results of our experiments across three sentiment analysis tasks: political sentiment analysis in Tamil and code-mixed sentiment analysis in Tamil-English and Tulu-English.

5.1 Political Sentiment Analysis

For the Tamil political sentiment classification task, our experiments with three different models showed that the Tamil-BERT model achieved the best performance with a macro F1 score of 0.36, outperforming both the multilingual BERT (0.27) and GPT-2 (0.26) models. On the final held-out test set, our system achieved a macro F1 score of 0.29, placing 9th in the competition rankings.

5.2 Code-Mixed Sentiment Analysis

The results for code-mixed tasks demonstrated notably different performance levels between Tulu-English and Tamil-English classification. Our system achieved exceptional performance on the Tulu-English task, with an overall accuracy of 0.92 and a macro F1 score of 0.84 on the validation set. This strong performance translated to the final evaluation, where our system ranked first in the competition for the Tulu-English track.

The Tamil-English task presented greater challenges, with our system achieving an accuracy of 0.45 and a macro F1 score of 0.19 on the validation set. The model showed stronger performance in identifying positive sentiments compared to other categories, but struggled with mixed feelings and unknown states.

5.3 Analysis

The disparity in performance between tasks can be attributed to several factors. The success in the Tulu-English task demonstrates the effectiveness of cross-lingual transfer learning, where a Telugu-BERT model successfully adapted to Tulu text. However, the Tamil-English task's lower performance highlights the challenges of handling larger

datasets with computational constraints. The political sentiment task's moderate performance reflects the inherent complexity of fine-grained sentiment classification in political discourse.

6 Conclusion

In this paper, we presented our approach to sentiment analysis across multiple Dravidian language tasks, including political sentiment classification in Tamil and code-mixed sentiment analysis in Tamil-English and Tulu-English. Our experiments demonstrated the effectiveness of language-specific transformer models, particularly in cross-lingual scenarios, achieving first place in the Tulu-English task using a Telugu-BERT model.

The varying performance across tasks—from high accuracy in Tulu-English to moderate results in political sentiment analysis—highlights both the potential and limitations of current approaches. Our findings suggest that while transformer-based models can effectively handle complex sentiment classification tasks, their success depends significantly on factors such as dataset characteristics and the specific nature of the sentiment analysis task. Future work could focus on developing specialized architectures for political discourse analysis and improving performance on larger-scale code-mixed datasets.

7 Limitations

Our work in Dravidian language sentiment analysis, while showing promising results, has several important limitations that warrant discussion. The reliance on language-specific BERT models, while effective, introduces significant computational constraints. The models require substantial GPU resources for training, particularly evident in the Tamil-English task with its larger dataset. The need for gradient accumulation in the Tamil-English model to manage memory constraints impacts training dynamics and potentially limits model performance. A critical limitation in our approach is the absence of a dedicated pre-trained model for Tulu, unlike the availability of specific models for Tamil and Telugu. While our cross-lingual transfer from Telugu to Tulu proved successful, having a Tulu-specific pre-trained model could have potentially captured more nuanced linguistic features and improved performance further.

The datasets present their own set of challenges, exhibiting substantial class imbalances particularly evident in the political sentiment task where the "None" category comprises only 3.9% of the data. Despite implementing label smoothing and weight decay to address this imbalance, the effectiveness of our models on minority classes remains limited. The social media origin of our data introduces additional complexity - social media text often contains informal language, abbreviations, and region-specific expressions that may not be well represented in our training data. Our code-mixed sentiment analysis, focusing specifically on Tamil-English and Tulu-English combinations, may not extend well to other code-mixing patterns or to texts with more than two languages mixed together, which is common in many Indian social media contexts. For instance, a single post might contain Tamil, English, and Hindi, but our current approach cannot handle such multi-language mixing effectively.

Political sentiment analysis presents unique challenges due to its dynamic nature. Our models don't explicitly account for temporal dynamics or evolving political contexts, which is particularly important in political discourse where the meaning and sentiment of certain terms or phrases can shift rapidly based on current events. The models might need regular retraining to maintain accuracy as political discourse and language usage patterns change over time. Furthermore, political sentiment often requires understanding subtle contextual cues, historical references, and cultural nuances that may not be fully captured by our current modeling approach. These limitations point to several directions for future research, including developing more efficient architectures for code-mixed text processing, creating dedicated pre-trained models for low-resource Dravidian languages, and designing approaches that can better handle the dynamic nature of political discourse.

References

Mohd Zeeshan Ansari, Mohd-Bilal Aziz, MO Siddiqui, H Mehra, and KP Singh. 2020. Analysis of political sentiment orientations on twitter. *Procedia computer science*, 167:1821–1828.

Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. 2013. Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020.

Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Tarek Elghazaly, Amal Mahmoud, and Hesham A Hefny. 2016. Political sentiment analysis using twitter data. In *Proceedings of the International Conference on Internet of things and Cloud Computing*, pages 1–5.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, Lavanya S K, Thenmozhi D., Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in Tamil and Tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Lavanya S. K., Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 2024. Overview of Second Shared Task on Sentiment Analysis in Code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

DLTCNITPY@DravidianLangTech 2025 Abusive Code-mixed Text Detection System Targeting Women for Tamil and Malayalam Languages using Deep Learning Technique

Habiba A , Aghila G

Department of Computer Science and Engineering,
National Institute of Technology Puducherry, India.

Abstract

The growing use of social communication platforms has seen women facing higher degrees of online violence than ever before. This paper presents how a deep learning abuse detection system can be applied to inappropriate text directed at women on social media. Because of the diversity of languages and the casual nature of online communication, coupled with the cultural diversity around the world, the detection of such content is often severely lacking. This research utilized Long Short-Term Memory (LSTM) for abuse text detection in Malayalam and Tamil languages. This model delivers 0.75, a high F1 score for Malayalam, and for Tamil, 0.72, achieving the desired balance of identifying abuse and non-abusive content and achieving high-performance rates. The designed model, based on the dataset provided in DravidianLangTech@NAACL2025 (shared task) comprising code-mixed abusive and non-abusive social media posts in Malayalam and Tamil, showcases a high propensity for detecting accuracy and indicates the likely success of deep learning-based models for abuse text detection in resource-constrained languages.

1 Introduction

The digital era has changed how people interact with each other and the rest of the world. Through social media, people can communicate with others who are oceans apart, share and experience events, and give opinions on matters openly. Social media has, however, advanced into a popular space that allows people to engage in passive yet harmful behaviours, most notably gendered abuse (De la Parra-Guerra et al., 2025). Some of the most common examples of these forms of abuse include the use of derogatory, threatening, or even demeaning language directed towards women and other targets (Gonzalez et al., 2025). This paper aims to comprehend the phenomena of aggressive Tamil and Malayalam language texts targeting women in social media by trying to find the patterns of the abuse and providing effective measures against them. Abusive language, particularly in social media, is communication intended to cause emotional or physical harm to others by using harmful or offensive verbal or textual content (Sinclair et al., 2025). Such verbal abuse may come in the form of name-calling, vulgar threats, and even sexual harassment. It is also aimed towards people who fall into specific categories, like women, ethnic minorities, or people from lower economic status. When this abuse is directed at women, it lacks compassion at its core, and this type of internet abuse is very damaging because it inflicts grave psychological, social and economic harm while reinforcing abuse of gender discrimination. The reason that understanding such abusive texts in Tamil and Malayalam is essential is due to a rise in online abuse within the South Indian linguistic paradigm. Among Indian languages, Tamil and Malayalam are among the most widely spoken. And their speakers form a vast and diverse population on social media. However, there have been no systematic studies on the use of abusive language in these languages and its impact on the dominant gender, particularly women. This understanding is necessary to form protective measures against unrelenting online abuse towards women in Tamil and Malayalam languages. In addition, these regional languages need to be studied to formulate more effective strategies focusing on prevention. The abuse is commonly used as a means to silence women, intimidate them, or disparage their voices in public debates, especially for women who speak about matters dealing with politics, gender, or social justice issues. For instance, women journalists, activists, and other women in powerful and visible public positions experience extreme forms of online abuse across all platforms, even in English, Spanish or Arabic. The verbal abuse is aimed at asserting supremacy over women and lowering

their self-esteem to the point where they are afraid to speak in public spheres (Albladi et al., 2025). There is ample literature that documents the issue of online abuse of women in different languages (Priyadharshini et al., 2022b, 2023b). Still, much less research has been done on how such abuse exists in Dravidian languages such as Tamil and Malayalam. The construction of identity is based on the language, and those regional languages that shape identity also come with cultural baggage that defines the abuse. Traditional gender norms, social discrimination, and particular dialects may determine the form that violence and abuse will take in Tamil and Malayalam. Hence, it deliberately illustrates the need to design an abusive text detection strategy for Malayalam and Tamil, incorporating informal language structures on various social media platforms.

2 Related Works

Existing models on abusive text detection have been successful in languages such as English, but their feasibility for languages like Malayalam and Tamil with less annotated data has been largely unexplored. These languages are uniquely different from English as both are Dravidian languages (Chakravarthi et al., 2021; Priyadharshini et al., 2023a), making the application of existing models impractical. Moreover, the challenge intensifies because large-scale annotated data for these languages are not available to the public. The informal use of language on social networks adds to the problem. For example, the frequent use of slang, abbreviations, code-mixing, and code-switching creates more problems for the existing text classification models. The following section outlines key findings from recent studies addressing this issue. Machine learning has been widely adopted to classify abusive language over the past few years. Support Vector machines (SVM), Random forests, and Naive Bayes classifiers have been used in (Mahmud et al., 2024; Thavareesan and Mahesan, 2019) to categorize abusive language using pre-defined features of the text. Nevertheless, a lot of them (Aljero and Dimililer, 2020; HaCohen-Kerner and Uzan, 2021) depend on excessive feature engineering, such as the formation of words or n-grams for hate speech detection. However, the emergence of deep learning has enormously impacted text classification processes and detecting abuses in text. RNNs and their later developments, Long Short-Term Memory (LSTM) networks, are used in (Zhang, 2024; Al-Qerem et al., 2024). Gated Recurrent Units (GRUs) are highly efficient in comprehending sequential information and perfect for text. These models can understand the relation between words and phrases in contexts, which helps identify abuses in context-sensitive language. More recent approaches, such as BERT, which is a transformer model, have been successful in (Tarun et al., 2024) over earlier methods of abusive text detection due to the model's ability to comprehend higher levels of representation semantics. The problem of abusive text detection in languages with relatively little data is complex. The presence of low-quality annotated datasets for languages such as Malayalam and Tamil makes it harder for models to be trained and to attain performance standards. In addition, the informal language that includes slang, code-mixing, code-switching, and dialect makes it even more challenging to detect abuse consistently and is also highlighted in (Priyadharshini et al., 2022a; Subramanian et al., 2024). Efforts carried out in languages with limited resources have shown that there is a need to develop a specific approach for such languages that considers a language's unique features.

3 Methodology

This section describes the detection of abusive text in the Malayalam and Tamil languages using an LSTM. The methodology comprises the following steps: dataset gathering, feature extraction, model training, and evaluation. Figure1 illustrates the general pipeline used for the text detection system. The user-generated content from YouTube, a social media platform where women are abused, is the source from which a dataset of abusive and non-abusive text is compiled. This data set is gathered from (Rajiakodi et al., 2025), the shared task DravidianLangTech@NAACL2025, and contains text samples in Malayalam and Tamil. The steps utilized for the detection system are elaborated in Algorithm 1.

3.1 Datasets

Here, we focus on two specifically code-mixed datasets: Malayalam code-mixed, which contains text that has both Malayalam and English words, and Tamil code-mixed text, which contains text that has both Tamil and English. Here, we focus on two specifically code-mixed datasets: Malayalam code-

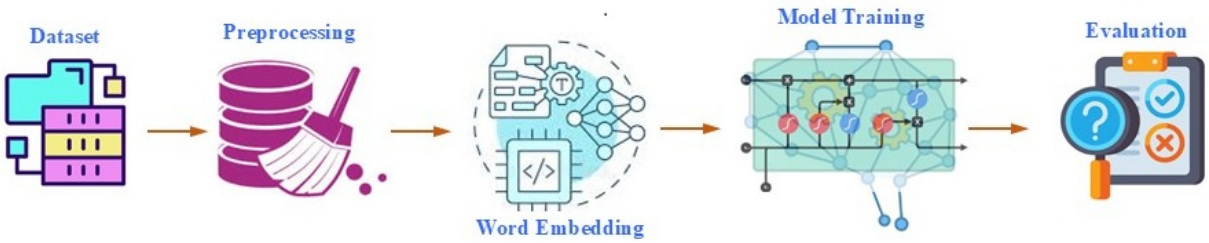


Figure 1: Text Detection System Pipeline

mixed, which contains text that has both Malayalam and English words, and Tamil code-mixed text, which contains text that has both Tamil and English. Table 1 describes the detailed statistics of both datasets. Both include text data categorized as abusive or non-abusive.

	Training Data		Validation Data	
	Abusive	Non-Abusive	Abusive	Non-Abusive
T	1366 (49%)	1424 (51%)	278 (46%)	320 (54%)
M	1531 (52%)	1402 (48%)	303 (48%)	326 (52%)

Table 1: Statistics of the dataset used for Tamil (T) and Malayalam (M).

3.2 Preprocessing

Before being used for analysis, text data goes through various cleaning and standardization steps, an essential step in almost every Natural Language Processing task. First, all text is translated into lowercase format. While this addresses the need for the first essential step for cleaning and standardizing text data, it also ensures no capitalization issues occur. Then, the unwanted spaces are removed from the dataset. As a result of checking whether the data has been prepared adequately for analysis, it serves as the basis for developing an efficient text classification model. Text data in the model undergoes tokenization, transforming each word into a unique number. The models scarcely use an extensive vocabulary and don't entertain less common words. The tokenized text is also padded so that all input data has the same shape.This preprocessed data is fed into the model for training.

3.3 Our Approach

The model used in this study is an LSTM, a specific RNN structure capable of working with sequences like string text. An architecture of this type includes an embedding layer, LSTM layer, and dense output layer. Word indices are mapped to fixed-sized dense vectors by the embedding layer. In contrast, the spatial dropout layer, which is set over the embedding layer, decreases overfitting for the model by setting a proportion of the embedding inputs to blank. The LSTM layer maintains the sequential relationships among the words. To avoid overfitting, dropout is used with both the input and recurrent connections. The dense output layer has two units enhanced with a sigmoid activation function for 2 class target variables. The model is compiled with Adam as the optimiser and binary cross-entropy loss, performing well for binary classification problems. During training, the model accuracy throughout evaluation is maintained. The model is trained for five epochs with a batch size of 64. The hyperparameter settings used to train the model are shown in the table 2. The validation data assesses the model's performance on the new test datasets that the model has not seen after every epoch. The metric F1-score is used to check the model's performance.

Parameters	Value
Embed Units	EMBEDDING_DIM = 100
Hidden Units (LSTM)	100
Dropout	0.2
Optimizer	Adam
Batch Size	64
Loss	Binary Crossentropy
Epochs	5
Activation	Sigmoid

Table 2: Hyperparameter Table

Algorithm 1 Text Classification for Abusive Language Detection using LSTM

Input: Dataset D with Text Sequences S

Output: Categories [$Abusive$ or $Non-Abusive$].

The input sequences are tokenized.

$$Tokenized(S) = [t_1, t_2, t_3, ..., t_n]$$

where t_i is the token for word w_i.

Using an embedding layer, each token t_i is mapped to a dense vector representation.

$$Embedding(t_i) = (e_{i1}, e_{i2}, e_{i3}, ..., e_{id})$$

where e_i represents the embedding vector for token t_i.

The embeddings are passed through the LSTM, and the output is a sequence of hidden state vectors:

$$h_i = LSTM(e_i, h_{i-1})$$

where h_i is the hidden state at time step i, and h_{i-1} is the previous hidden state.

The final hidden state h_n is used for classification.

The dense layer produces a vector y_{logits} for the final prediction:

$$y_{logits} = Wh_n$$

where W is the weight matrix and h_n is the final hidden state.

The output probabilities are calculated:

$$p(y) = [p(0), p(1)]$$

where $p(0)$ is the "Non-Abusive" class probability, and $p(1)$ is the "Abusive" class probability.

The binary cross-entropy loss function is used to measure the discrepancy between the predicted probabilities:

$$BC\ Loss = -\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i) \log(1-p_i)$$

where N is the number of samples in the batch, y_i is the actual label for the i-th sample, and p_i is the predicted probability for the i-th sample.

4 Results

The model improved the accuracy of the Tamil code mixed dataset throughout the training epochs. The accuracy result is 78.48%, and the F1 score is 0.7207 for the Tamil test set. The model placed 18^{th}, with an impressive quantitative score. The model also improved significantly during the training epochs for the Malayalam code mixed dataset. The accuracy achieved was 73.50, and an F1 score of 0.7571 from the Malayalam dataset, showing balanced classification results. It is impressive that the model ranked 1 for Malayalam in the shared task.The table 3 shows the results achieved through our approach. The model handled the Tamil and Malayalam datasets well, making it the best model for this category.

Team Name	Language	F1	Rank
Habiba A, Aghila G (This work)	Malayalam	0.7571	1
	Tamil	0.7207	18

Table 3: Result Achieved

5 Discussions

Although our model performed well, we recognize the weaknesses of the LSTM architecture. In future, we have a strategy to resolve developmental possibilities, such as looking into more complex neural networks, incorporating more factors such as sentiment, and taking advantage of more significant and varied datasets.

6 Conclusion

This paper outlines a methodology wherein deep neural networks detect abusive texts in multiple languages, especially those that lack sufficient training data, like Malayalam and Tamil. Using RNNs and their ability to detect patterns, we aim to create an accurate model for abuse text detection in women. The findings, as appreciated, need an elaboration on concepts like an informal conversation, the use of idioms, and the absence of sufficient labelled datasets. The model is trained for each particular language separately to accommodate their diverse alterations in abusive language.

7 Error Analysis

While the LSTM model used on Tamil and Malayalam hate speech shows excellent accuracy, espe-

cially with true positives, it faces challenges with false positives and other classed errors within a supposedly balanced dataset. This particular behaviour needs much attention and improvement on the model's discrimination capabilities. Evaluation of the model's performance through comprehensive validation and test set analyses is vital for determining performance generalizability. Some of the outlined tactical changes include changing the degree of false positive identification and incorporating large language models for optimization.

8 Limitations

The pertinent issues of this analysis are the inadequacy of the annotated dataset for both Malayalam and Tamil—more significant datasets aid model effectiveness and generalizability. The preprocessing and classification of such texts are complicated due to the lack of their formal quality. And indeed, language, as used in social media, constructed with informal terms, regional variations, and shortened expressions, is highly challenging. Most of these phrases lack grammatical structure, which hinders traditional models from determining the sentiment of the text. Both these languages, Malayalam and Tamil, are classified under the same Dravidian language family yet differ in morphology, syntax, semantics and other unique factors.

Ethics Statement

We maintain compliance with ACL guidelines in participating DravidianLangTech@NAACL2025 shared task, as well as our commitment to ethical research practices. No ethical issues or conflicts of interest emerged throughout the duration of this research.

References

Ahmad Al-Qerem, Mohammed Raja, Sameh Taqatqa, and Mutaz Rsmi Abu Sara. 2024. Utilizing deep learning models (rnn, lstm, cnn-lstm, and bi-lstm) for arabic text classification. In *Artificial Intelligence-Augmented Digital Twins: Transforming Industrial Operations for Innovation and Sustainability*, pages 287–301. Springer.

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892.

Mona Khalifa A. Aljero and Nazife Dimililer. 2020. Hate speech detection using genetic programming. *2020 International Conference on Advanced Science and Engineering (ICOASE)*, pages 1–5.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. Mccrae. 2021. Dravidiancodemix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56:765 – 806.

AC De la Parra-Guerra, J Truyol-Padilla, CA García-Alzate, and F Fuentes-Gandara. 2025. Gender-based violence as a barrier to women rights towards socio-environmental sustainability. *Global Journal of Environmental Science and Management*, 11(1):343–364.

Alejandra Gonzalez, James K Haws, Nuha Alshabani, Caron Zlotnick, and Dawn M Johnson. 2025. Cyber abuse and posttraumatic stress disorder among racially diverse women who have resided in domestic violence shelters: A longitudinal approach. *Psychological Trauma: Theory, Research, Practice, and Policy*.

Yaakov HaCohen-Kerner and Moshe Uzan. 2021. Detecting offensive language in english hindi and marathi using classical supervised machine learning methods and word/char n-grams. In *FIRE (Working Notes)*, pages 501–507.

Tanjim Mahmud, Tahmina Akter, Mohammad Kamal Uddin, Mohammad Tarek Aziz, Mohammad Shahadat Hossain, and Karl Andersson. 2024. Machine learning techniques for identifying child abusive texts in online platforms. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022a. Overview of abusive comment detection in tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022b. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, S Malliga, Subalalitha Cn, SV Kogilavani, B Premjith, Abirami Murugappan, and Prasanna Kumar

Kumaresan. 2023a. Overview of shared-task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023b. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Gary Sinclair, Colm Kearns, Katie Liston, Daniel Kilvington, Jack Black, Mark Doidge, Thomas Fletcher, and Theo Lynn. 2025. Online abuse, emotion work and sports journalism. *Journalism Studies*, 26(1):101–119.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

VG Tarun, Ramkumar Sivasakthivel, Gobinath Ramar, Manikandan Rajagopal, and G Sivaraman. 2024. Exploring bert and bi-lstm for toxic comment classification: A comparative analysis. In *2024 Second International Conference on Data Science and Information System (ICDSIS)*, pages 1–6. IEEE.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on industrial and information systems (ICIIS)*, pages 320–325. IEEE.

Hongmin Zhang. 2024. Research on text classification based on lstm-cnn. In *Proceeding of the 2024 5th International Conference on Computer Science and Management Technology*, pages 277–282.

Hydrangea@DravidianLanTech2025: Abusive language Identification from Tamil and Malayalam Text using Transformer Models

Shanmitha Thirumoorthy
Vellore Institute of Technology
shanmitha.t2023@vitstudent.ac.in

Ratnavel Rajalakshmi
Vellore Institute of Technology
rajalakshmi.r@vit.ac.in

Durairaj Thenmozhi
Sri Sivasubramaniya Nadar College of Engineering
theni_d@ssn.edu.in

Abstract

Abusive language toward women on the Internet has always been perceived as a danger to free speech and safe online spaces. In this paper, we discuss three transformer-based models - BERT, XLM-RoBERTa, and DistilBERT-in identifying gender-abusive comments in Tamil and Malayalam YouTube contents. We fine-tune and compare these models using a dataset provided by DravidianLangTech 2025 shared task for identifying the abusive content from social media. Compared to the models above, the results of XLM-RoBERTa are better and reached F1 scores of 0.7708 for Tamil and 0.6876 for Malayalam. BERT followed with scores of 0.7658 (Tamil) and 0.6671 (Malayalam). Of the DistilBERTs, performance was varyingly different for the different languages. A large difference in performance between the models, especially in the case of Malayalam, indicates that working in low-resource languages is difficult. The choice of a model is extremely critical in applying abusive language detection. The findings would be important information for effective content moderation systems in linguistically diverse contexts. In general, it would promote safe online spaces for women in South Indian language communities.

1 Introduction

The digital revolution has transformed social networks from a double-edged sword where it fosters democratization of communication, but also allows systemic gender-based violence through abusive content targeting women. Above 40% of the women worldwide said they had suffered online harassment; therefore, South Asian contexts are generally more vulnerable, mainly due to linguistic complexity and domination patriarchal norms. This abuse comes in the form of overt threats, veiled misogyny, and damaging stereotypes that lead to psychological trauma in the form of depression and anxiety, professional reversals, and even physical danger, which bleed into the offline world. Low-resource Dravidian languages like Tamil and Malayalam pose a heightened challenge because of morphological complexity, patterns of code-mixing, and scarce NLP resources: South Asian languages are spoken by 300 million people, but only 0.1% of AI research focuses on these languages. Advanced text classification techniques are used in automated content moderation systems, which are scalable. Especially promising for unmasking contextual abuse patterns is the transformer model. We can take the "digital silencing" effect stemming from the fact that 30% of women journalists self-censor due to online threats and develop language-specific detection frameworks, therefore cultivating safer digital ecosystems that empower rather than endanger women.

DravidianLanTech shared task (Priyadharshini et al., 2022) (Priyadharshini et al., 2023) is continuously focusing on the abusive language identification in the social media contents in Dravidian languages. DravidianLangTech 2025 (Rajiakodi et al., 2025) gives emphasis to Tamil and Malayalam languages in identifying the language targeting to women in social media. Our team Hydrangea participated in this shared task and submitted three runs for both languages. The code is found in GitHub Link

2 Related Works

Several research works have been carried out for detecting abusive content in social media. They used from traditional classifiers to transformer models for finding the same.

(Bansal et al., 2022) used XLM RoBERTa, a model pre-trained on more than 100 languages, and added a BiGRU layer for the classification of abusive content in 13 low-resource languages, including Hindi, Telugu, Marathi, and Tamil. Due to a robust pre-training and handling capacity of

the model, it was feasible to handle datasets with less availability of natural language processing resources that exceeded traditional methods like Naive Bayes and Logistic Regression. The study was conducted on four different transformer models namely mBERT, MurilBERT, IndicBERT, and XLM RoBERTa along with several data processing techniques like cleaning and transliteration with emoji embeddings. XLM RoBERTa performed the best where it shows superiority over language-specific models with excellent performance and efficiency. In general, the results showed that the importance of using transformer-based approaches as well as data preprocessing to improve the accuracy of models in performing multilingual tasks.

(Philipo et al., 2024) presented the evaluation of three large language models, BERT, XLM RoBERTa, and DistilBERT, as candidates for using an annotated standardized corpus of texts in recognizing cyberbullying on social media. Fine-tuned each model in such a manner that the trained models now worked as two-class classifiers - bullying or not bullying-the texts, and with optimised usage of resources along with good accuracy during classification. Key metrics was then measured on the parameters such as accuracy, precision, recall, and F1-score.

The best model was BERT in all the key metrics, performing at 95%; it seems to recognize subtle language motifs, suggestive of bullying. Both precision and recall metrics indicate excellent capacity for correctly classifying bullying as well as non-bullying instances-the model is showing good balance. XLM RoBERTa showed good performance but was less efficient compared with BERT. This accounted for slightly lower metrics overall because the dataset was monolingual. DistilBERT was designed as a lighter version of BERT and showed increased computational efficiency with faster inference times but recorded slightly lower accuracy and F1-scores. Thus BERT emerged victorious in all metrics

(Koufakou et al., 2020) evaluated the in-domain and cross-domain performance of the basic BERT model against enhanced HurtBERT based on an extensive experimental setup. In the in-domain tasks, where training and test data have similar distributions, the additional lexical features from a hate lexicon are proved to be quite useful for HurtBERT: enhancing precision, recall, and F1-score by identifying explicit abusive language patterns often missed by BERT's contextual embeddings. For cross-domain tasks, where training was done on one dataset and testing on another, HurtBERT was better in generalization because it used sentence-level lexicon encodings and word-level embeddings to adapt effectively to different linguistic styles and representations of abusive content. The problem with transformer-based models was not solved, or rather, in domains with poor amounts of labelled data or differing text styles. The above discussions reflect the opportunity of HurtBERT in practical scenarios of abusive language detection.

(Manikandan et al., 2022) used a three-stage system architecture that included pre-processing, model training, and testing using two transformer models, BERT and XLM-RoBERTa. Despite the fact that both of these models were trained on preprocessed data, in all of these metrics, XLM-RoBERTa surpassed BERT especially with homophobic content, which gave 93% accuracy against 91% BERT, the results reflected that XLM-RoBERTa works well with homophobic content and both faced the problem while handling the few samples of transphobic content.

(Gayathri et al., 2022) used support vector machine with LaBSE embedding for finding the abusive language in Tamil text. (Gayathri et al., 2024) employed a combination of statistic features and language-agnostic features and performed feature selection by using explainable AI for detecting abusive language in Tamil-English codemixed text.

3 Data Description

The organizer released 2 sets of dataset. Tamil data set consists of 2790 instances in the training set in which 1366 instances are Abusive category and 1424 are Non-abusive category. Malayalam training data set has a total of 2935 instances with 1531 Abusive contents and 1402 Non-abusive contents. The test data set has 598 and 629 instances for Tamil and Malayalam respectively. The data distribution shows that the data set is balanced and is not required much of balancing the data.

4 Methodology

The DravidianLangTech 2025 abusive language identification came equipped with a 4,543 annotated comment data set - 2,819 Tamil and 1,724 Malayalam. The data set included YouTube comments labeled with binary tags: "Abusive" and "Non-Abusive". Tamil and Malayalam are low-resource languages, and it is difficult to work

with them in the area of natural language processing. The data was gathered from YouTube comments and pre-cleaned to eliminate unwanted characters and normalize the text. The abusive and non-abusive comment distribution was not entirely balanced, but slightly more comments were non-abusive. We employed three of the most used architectures of text classification tasks for this abusive language identification task in Tamil and Malayalam: BERT, XLM-RoBERTa, and DistilBERT. Datasets compatible with each other were utilized to tune the models for classifying the comments as appropriate.

BERT is essentially a pre-trained model based on roughly 2.5 billion words taken from Wikipedia in 104 languages using a vocabulary of approximately 110,000 word pieces. BERT functions with its principal training being MLM to better handle the bidirectional words relationship. As compared to traditional RNN, it does not see one token at a time but, rather, BERT randomly masks 15% of the input tokens in every layer and predicts masked tokens based on the entire input sequence. AdamW is employed for fine-tuning optimization and the learning rate is set at $2\mathrm{e}-5$.

In DistilBERT, architectural compression lowers the computational requirements of it. Knowledge distillation is employed to train a 6-layer model that mimics the output of BERT's 12-layer model.

Better performance by XLM-RoBERTa was due to pretraining on multiple languages across different linguistic structures, thus dynamically adapting the vocabulary to suit agglutinative morphological features of Tamil and Malayalam. Training was done with batch size 16, AdamW optimizer with a learning rate of $2\mathrm{e}-5$, and early stopping on validation loss.

For pre-training, BERT and XLM-RoBERTa were pre-trained for two epochs whereas DistilBERT was trained for three epochs because it has a smaller model and hence quicker training cycles.

5 Result and analysis

We have evaluated the three models on the datasets provided by the DravidianLangTech 2025 shared task. Tables 1 and 2 show the performance of our three models in terms of precision, recall and F1 score on the Tamil and Malayalam development data sets respectively. XLM-Roberta performed better for Tamil data set whereas BERT performed better for Malayalam data set.

Similarly for test data sets, 3 and 4 show the performance of our three models in Tamil and Malayalam respectively, where XLM-RoBERTa proved successful for both language data sets. XLM-RoBERTa showed the highest classification performance in F1-score at 0.7708 (Tamil) and 0.6876 (Malayalam) compared to BERT at 0.7658 Tamil and 0.6671 Malayalam and DistilBERT at 0.7639 Tamil and 0.4876 Malayalam. The models are very consistent with Tamil classification; all architectures showed F1-score above 0.76 while there is significant performance degradation for Malayalam, especially for DistilBERT near-random performance of 0.4876 . While BERT employed bidirectional attention to identify contextual patterns of abuse, its efficacy was lower on Malayalam with lesser pretraining exposure.

The performance difference between XLM-RoBERTa and BERT for Tamil and Malayalam may be due to various reasons. XLM-RoBERTa is particularly good at multilingual pre-training, where it is able to identify cross-lingual similarities and transfer knowledge across languages. Tamil may have been able to learn more from the cross-lingual knowledge transfer because of its linguistic proximity with other languages in the pre-training data, or being better represented in that data. On the other hand, the distinct linguistic features of Malayalam or its under-representation within the pre-training data could have impeded XLM-RoBERTa's performance in relation to BERT, which perhaps was more well-suited to pick up the particular idiosyncrasies of the Malayalam dataset.

Table 1: Performance Comparison on Tamil-English Development Set

Model	Precision	Recall	F1
BERT	0.5000	0.5000	0.5000
XLM-RoBERTa	0.2900	0.5400	0.5400
DistilBERT	0.4800	0.4800	0.4700

Table 2: Performance Comparison on Malayalam-English Development Set

Model	Precision	Recall	F1
BERT	0.5000	0.5000	0.5000
XLM-RoBERTa	0.2300	0.4800	0.4800
DistilBERT	0.4500	0.4700	0.4600

The F1-scores of the test case are low, and they can inform us where something went awry while

Table 3: Performance Comparison on Tamil-English Test Set

Model	Precision	Recall	F1
BERT	0.7658	0.7658	0.7658
XLM-RoBERTa	0.7708	0.7708	0.7708
DistilBERT	0.7639	0.6307	0.6909

Table 4: Performance Comparison on Malayalam-English Test Set

Model	Precision	Recall	F1
BERT	0.6671	0.6613	0.6641
XLM-RoBERTa	0.6817	0.6722	0.6769
DistilBERT	0.4798	0.4026	0.4378

pre-processing data and training the model. DistilBERT did not perform for Malayalam, and it may also be a matter of its smaller size not being able to keep up with the complexity of the language or of the model requiring larger training data sets due to this reason. There are further experiments to be conducted in order to ascertain the reason behind the performance of DistilBERT having a negative effect on Malayalam. There may be a reason that has something to do with too little pre-training of DistilBERT on Malayalam text data, and this implies Malayalam-specific linguistic features are under-represented in the model parameters. There may be a second reason, though, and this has something to do with the case of low-quality Malayalam data sets being used to train DistilBERT, i.e., abusive language data sets which do not allow DistilBERT to learn discriminatory patterns successfully.

6 Limitation

The transformer models used in our experiments lack in finding patterns from code-mixed Dravidian languages, resulting in low F1 scores. This is mainly due to the size of the data set which is not adequate to train the transformer model. This may be overcome by applying some data augmentation techniques in future.

7 Conclusions

Transformer models i.e. BERT, XLM-RoBERTa and DistilBERT are employed to detect the abusive content targeting women on social media. Our team Hydrangea achieved 9th and 12th ranks with macro F1 scores of 0.7708 and 0.6769 for Tamil and Malayalam, respectively in the leader board utilizing the models trained on XLM-RoBERTa. Language-agnostics embeddings can be utilized in the future to enhance the performance of our approach. In addition, Explainable AI (XAI) methods like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) can be applied to discover how the model is making its decision by determining the words or phrases most responsible for abusive classification. Attention visualization approaches may also be utilized, especially for transformer models, to identify what aspects of the input text the model is paying attention to. These observations can then be utilized to tune feature weights and tune the model's structure, possibly leading to enhanced performance. Zero-shot or few-shot learning methods may also be employed to surmount the problem of data sparsity, by utilizing knowledge from other languages or tasks to enhance performance on Tamil and Malayalam with little training data.

References

Vibhuti Bansal, Mrinal Tyagi, Rajesh Sharma, Vedika Gupta, and Qin Xin. 2022. A transformer based approach for abuse detection in code mixed indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

G L Gayathri, Krithika Swaminathan, Divyasri Krishnakumar, Thenmozhi D, and Bharathi B. 2024. Abusive comment detection in tamil code-mixed data by adjusting class weights and refining features. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

GL Gayathri, Krithika Swaminathan, K Divyasri, Thenmozhi Durairaj, and B Bharathi. 2022. Pandas@ abusive comment detection in tamil code-mixed data using custom embeddings with labse. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 112–119.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Deepalakshmi Manikandan, Malliga Subramanian, and Kogilavani Shanmugavadivel. 2022. A system for detecting abusive contents against lgbt community using deep learning based transformer models. In *FIRE (Working Notes)*, pages 106–116.

Adamu Gaston Philipo, Doreen Sebastian Sarwatt, Jianguo Ding, Mahmoud Daneshmand, and Huansheng Ning. 2024. Assessing text classification methods for

cyberbullying detection on social media platforms. *Preprint*, arXiv:2412.19928.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith , Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the Shared Task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

CUET_NLP_FiniteInfinity@DravidianLangTech 2025: Exploring Large Language Models for AI-Generated Product Review Classification in Malayalam

Md. Zahid Hasan, Safiul Alam Sarker, MD Musa Kalimullah Ratul
Kawsar Ahmed and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904099, u1904041, u1904071, u1804017}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Detecting AI-generated product reviews is a critical challenge in natural language processing (NLP), particularly for low-resource languages like Malayalam. In this study, we propose a large language model (LLM)-based approach to identify AI-generated reviews in Dravidian languages, focusing on the product review domain. We systematically evaluated multiple LLMs on a dedicated Malayalam dataset to assess their effectiveness in distinguishing between human-written and AI-generated reviews. Our experiments demonstrate that the Gemma-2B model outperforms other models, achieving a macro F1-score of 89.99%.Our approach secured 5^{th} place in the DravidianLangTech@NAACL 2025 shared task for Malayalam, highlighting the potential of LLMs in tackling the challenges of AI-generated review detection in low-resource languages. Our findings highlight the potential of LLMs in detecting AI-generated content in underrepresented languages, contributing to advancements in Dravidian language processing and the broader field of AI-generated content identification.

1 Introduction

The detection of AI-generated product reviews is an emerging challenge in natural language processing (NLP), particularly in the context of low-resource languages. As AI-generated text becomes increasingly sophisticated, distinguishing between human-written and machine-generated content is crucial for ensuring the authenticity of online reviews. In e-commerce platforms, product reviews play a pivotal role in shaping consumer trust, influencing purchasing decisions, and maintaining credibility between buyers and sellers. The proliferation of AI-generated reviews, however, poses a significant risk to the reliability of these platforms, making automated detection systems essential.

While extensive research has been conducted on AI-generated text detection in English (Salminen et al., 2022; Luo et al., 2023), there remains a notable gap in studies focusing on particularly Malayalam language. The limited availability of high-quality annotated datasets, coupled with the linguistic complexity of these languages, presents significant challenges in building robust detection models. This shared task addresses these gaps by encouraging the development of effective machine learning approaches for detecting AI-generated product reviews in Malayalam arranged by DravidianLangTech@NAACL 2025 (Premjith et al., 2025).

In this study, AI-generated product review detection is formulated as a binary classification problem, where the goal is to classify a given review as either human-written or AI-generated. To tackle the challenges associated with low-resource languages, we explore the application of large language models (LLMs) fine-tuned specifically for this task. Our contributions are as follows:

- We propose a fine-tuned large language model designed for AI-generated product review detection in Malayalam language.
- We systematically evaluate multiple LLMs including Gemma-2-2b (Team et al., 2024), Llama-3.2-3B (AI, 2024), sarvam-1[1], Qwen2.5-3B (Yang et al., 2024), and BharatGPT-3B-Indic[2] to identify the most effective approach for this task.

This work aims to advance the field of AI-generated text detection in low-resource language and establish a strong foundation for future research in this domain.

[1] https://huggingface.co/sarvamai/sarvam-1
[2] https://huggingface.co/CoRover/BharatGPT-3B-Indic

2 Related Work

The growth of AI-generated content, especially in the form of product evaluations, has emerged as a serious concern in e-commerce and social media platforms. Recent research have emphasised the increased competence of AI models in creating human-like writing, making it more difficult to discern between genuine and false evaluations (Luo et al., 2023). Gambetti and Han (2023) suggested utilizing AI to combat machine-generated phony reviews, getting an F1 score of 0.92 on a restaurant review data set using ensemble learning and contextual embeddings. Similarly, Birim et al. (2022) applied topic modeling approaches to detect patterns suggestive of bogus reviews, reporting an accuracy of 89.5% on a multilingual dataset. Salminen et al. (2022) examined the development and detection of fake reviews, attaining an F1 score of 0.87 by integrating language characteristics and behavioral analysis. Shibani et al. (2024) examined generative AI for Tamil writing help, getting an F1 score of 0.89 in recognising AI-generated text. De et al. (2021) suggested a transformer-based technique for multilingual false news identification, obtaining 87.3% accuracy and an F1 score of 0.85 using mBERT. Budhi et al. (2021) handled unbalanced datasets in fake review identification, reaching 91.2% accuracy and an F1 score of 0.88 by resampling and textual characteristics. Cheng et al. (2024) used graph neural networks (GNNs) to detect bogus reviewers, attaining an F1 score of 0.92 by assessing social context. Mukherjee (2024) emphasises on avoiding AI-generated fraud in marketing, underlining the importance for robust detection techniques. These findings underscore the significance of language-specific and context-aware models, particularly for low-resource languages like Malayalam and indicate the promise of advanced techniques like transformers and GNNs in fighting AI-generated fraudulent information. These findings together underline the necessity for powerful, language-specific detection approaches, especially for low-resource languages like Dravidian languages, to face the rising threat of AI-generated bogus reviews.

3 Dataset and Task Description

The shared task on "Detecting AI-generated Product Reviews in Dravidian Languages" (Premjith et al., 2025) focuses on identifying AI-generated and human-written reviews in Malayalam. With the increasing sophistication of AI tools, this task addresses the need for accurate detection models in the domain of online reviews, where authenticity is critical.

Participants were provided datasets comprising human-written and AI-generated reviews. As shown in Table 1, the training set includes 800 reviews, while the test set contains 200 reviews. The task invites global participation via CodaLab[3] to enhance AI detection for Dravidian language.

Set	Class	S_C	W_T	W_U	Avg. Len
Train	HUMAN	400	6174	3357	15.44
	AI	400	4121	2317	10.30
Test	HUMAN	100	2027	1462	20.27
	AI	100	1053	821	10.53

Table 1: Dataset Statistics: Sentence Count (S_C), Total Words (W_T), Unique Words (W_U), and Average Length (Avg. Len)

4 System Overview

In this study, we investigate a comprehensive suite of approaches—including machine learning (ML), deep learning (DL), transformer-based methods, and large language models (LLMs)—to detect AI-generated product reviews in Malayalam. Figure 1 presents a schematic overview of our proposed methodology. Detailed implementation and source code for the system are available on GitHub[4].

4.1 Machine Learning Approaches

We evaluated traditional ML models, including Logistic Regression, Support Vector Machines (SVM), Random Forest (RF), Naïve Bayes (NB), Decision Trees (DT), Kernel SVM, and Stochastic Gradient Descent (SGD), for product review classification. Textual data was transformed into high-dimensional vectors using TF-IDF (Takenobu, 1994) and CountVectorizer, with TF-IDF limited to 1000 features for optimal performance. Logistic regression used a maximum of 1000 iterations. Both linear and kernel SVMs were tested, with the RBF kernel applied for non-linear classification and regularization parameter $C = 0.80$ for linear SVM.

4.2 Deep Learning Approaches

We explored several deep learning models, including CNN, BiLSTM, BiLSTM+Attention,

[3] https://codalab.org/

[4] https://github.com/zahid99hasan/AI-Generated-Text-Detection

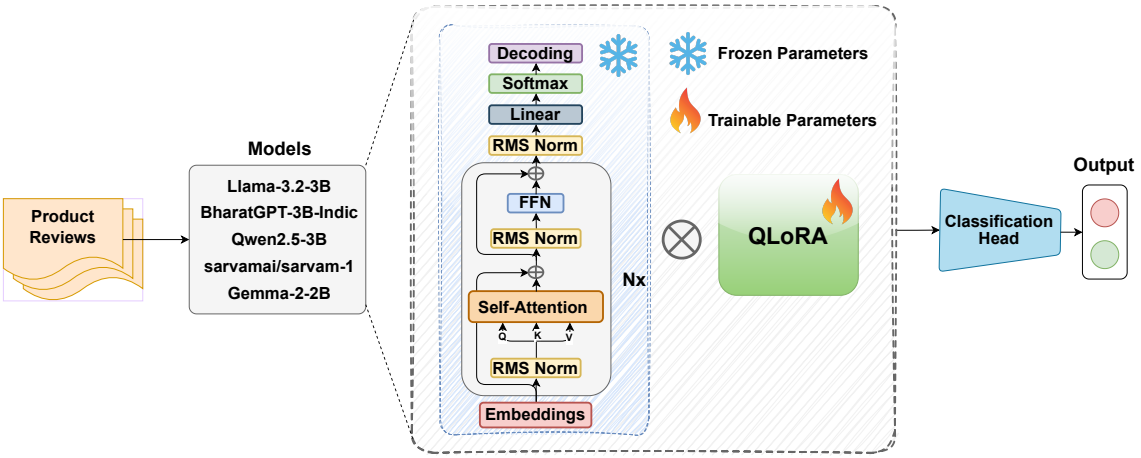


Figure 1: Schematic process for product review detection.

Keras+CNN, GloVe+CNN, and GloVe+BiLSTM. The CNN model utilized 128 filters with a kernel size of 5 and 24 neurons in the dense layer. In the BiLSTM+Attention model, ReLU was used in the dense layer and Sigmoid in the output layer. BiLSTM (Schuster and Paliwal, 1997) enhanced performance by capturing bidirectional contextual information. Keras+CNN and GloVe+CNN employed 1D convolutional layers for processing sequential data. GloVe+BiLSTM used an LSTM architecture with 64 neurons to effectively capture long-term dependencies in text sequences.

4.3 Transformer-Based Approaches

Several pre-trained transformer models from Hugging Face were leveraged for product review classification, including mBERT (Devlin, 2018), XLM-R (Conneau, 2019), MalayalamBERT (Joshi, 2022), and IndicBERT (Kakwani et al., 2020). Before passing data through the transformers, preprocessing and tokenization were performed. All models were trained using a learning rate of 5×10^{-5}, a batch size of 16 for both training and validation, and 4 epochs to achieve optimal results.

4.4 LLM-Based Approaches

Large language models (LLMs) with efficient fine-tuning via QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2024) were utilized for classifying product reviews in Dravidian languages (e.g., Malayalam) as Human-generated or AI-generated. Pre-trained LLMs, including Llama-3.2-3B (AI, 2024), BharatGPT-3B-Indic[5], Qwen2.5-3B (Yang et al., 2024), sarvamai/sarvam-1[6], and Gemma-2-2B (Team et al., 2024), were employed. QLoRA preserves base model parameters while introducing trainable low-rank adapters, enabling efficient task-specific adaptation. Figure 1 illustrates the proposed approach. The classification pipeline begins with tokenizing input reviews, which are then processed through frozen LLM layers to generate contextual embeddings. These embeddings are modified by QLoRA adapters, which introduce trainable low-rank updates while keeping the base model parameters unchanged. The adapted embeddings are then passed to a classification head, which performs the final prediction to determine whether the review is Human-generated or AI-generated. This approach efficiently adapts pre-trained LLMs for the classification task while maintaining computational efficiency and robust performance. The proposed model (Gemma-2-2B) was trained for 10 epochs with a batch size 16 and a learning rate of 1e-4, achieving the best overall performance. QLoRA was applied with a rank of 4, alpha of 16, a dropout rate of 0.1, and no bias.

5 Results and Analysis

Table 2 demonstrates the evaluation results large language models on the test set.

Results revealed that Gemma-2-2B for Malayalam earned the most elevated macro F1-score (89.99%) among the LLM approaches. On the other hand, for Malayalam sarvam-1 with macro F1-score (84.47%) surpasses all the models except Gemma-2-2B among the models. For Malay-

[5] https://huggingface.co/CoRover/BharatGPT-3B-Indic

[6] https://huggingface.co/sarvamai/sarvam-1

alam, Llama-3.2-3B perform poorly with the lowest macro F1 score.

ML Models			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
LoR	64.00	66.00	67.00
SVM	63.00	66.00	67.00
RF	61.00	65.00	65.00
NB	59.00	62.00	62.00
DT	59.00	57.00	57.00
KerneL SVM	63.00	66.00	67.00
SGD	**67.00**	**69.00**	**69.00**
DL Models			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
CNN	72.74	73.20	73.50
BiLSTM	64.99	66.66	68.00
BiLSTM + Attention	17.32	36.58	51.50
Keras + CNN	73.99	73.99	74.00
GloVe + CNN	66.11	68.32	70.00
GloVe + BiLSTM	**74.00**	**74.00**	**74.00**
Transformers			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
mBERT	**74.90**	**78.62**	**78.75**
IndicBERT	64.95	74.08	75.00
XLM-R	54.56	63.16	64.37
MalayalamBERT	70.57	77.66	78.12
LLMs			
Classifier	**G-mean(%)**	**F1(%)**	**Ac(%)**
Gemma-2-2B	**90.06**	**89.99**	**90.00**
sarvam-1	84.63	84.47	84.50
Llama-3.2-3B	35.35	33.33	50.00
Qwen2.5-3B	25.97	37.60	40.50
BharatGPT-3B-Indic	30.87	45.50	29.60

Table 2: Performance of the different methods on the test set

LLMs perform well on the validation set, which is splinted from train set. Over fitting can be a reason for that. BhartGPT-3B-Indic expected to perform really well, but in reality it shows an average performance. This encourage to explore more models for better performance. To explore better performance, this work explored Qwen2.5-3B, Llama-3.2-3B models as well.

5.1 Error Analysis

A comprehensive error analysis is performed to offer in-depth insights into the performance of the proposed model.

Quantitative Analysis

Since the gold labels for the test set were disclosed, Figure 2 presents the confusion matrix that categorizes product reviews into Human and AI predicted. The figure indicates that out of 200 reviews , 180 were accurately predicted. Al reviews (13) predicted more incorrectly compare to Human reviews (7), this occurred because AI can generate humanoid reviews. And for short-length data (less than or equal to 10 words), our proposed model perform better than large length of data. The proposed models only trained on the given dataset.

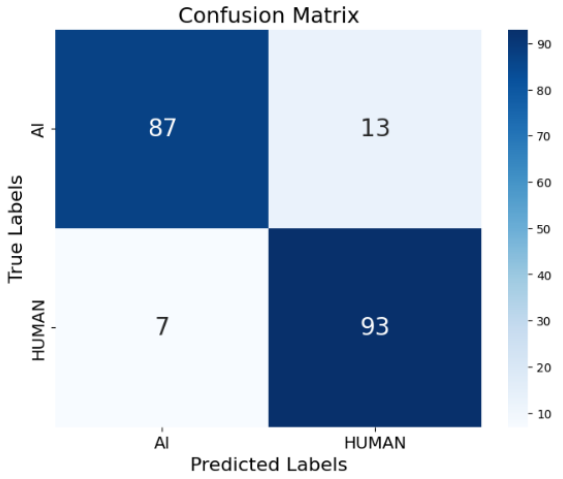


Figure 2: Confusion matrix of Gemma-2-2B model.

Qualitative Analysis

Table 3 presents some predicted outputs of the tested model. In the first, second and fourth data, the model successfully predicted the review of the data. On the other hand, it failed to do so in the third data. The Gemma-2-2B model for Malayalam, which is fine-tuned in this work, is primarily trained on given datasets from different sources; this could be one of the reasons for these model's failure in some Malayalam data. Furthermore, incorrect predictions arose due to stylistic similarities between formal human-authored text and AI-generated patterns, insufficient diversity in the training dataset, topical overlap with AI-prevalent themes, limitations in model capacity from low-rank adaptation and quantization, and loss of context from input truncation.

Text Sample	Actual	Predicted
Sample1: കോവയ്ക്ക ഉപ്പിലിട്ടത് എൻറെ ജീവിതത്തിൽ ഇതുവരെ കഴിച്ചിട്ടില്ല (I have never eaten salted cod in my life.)	Human	Human
Sample2: കോവയ്ക്ക ഉപ്പിലിട്ടത് ഞാൻ ഇതുവരെ കഴിച്ചിട്ടില്ല, കഴിക്കാൻ മനസില്ല. (I have never eaten salted cod, and I don't feel like eating it.)	AI	AI
Sample3: ഇക്കാലത്ത് ഒക്കെ ടാറ്റയുടെ ഡിസൈൻ കണ്ട് കണ്ണിന്ന ഊട്ടുപോലും ഇട്ടുന്നണ്ട് (Nowadays, Tata's designs are even making my eyes water.)	AI	Human
Sample4: കാൽ കാശിന്ന കൊള്ളാത്ത ഭക്ഷണമാണ് ചേട്ടാ (The food is not worth the money, brother.)	Human	Human

Figure 3: Sample predictions with actual and predicted reviews

6 Conclusion

This study explored the effectiveness of several large language models in detecting AI-generated content within a Malayalam product review dataset. Our findings indicate that the Gemma-2-2B model excelled in this task, achieving a macro F1-score of 89.99%. The results underscore the potential of transformer-based approaches for this application and motivate further exploration of alternative transformer architectures and LLMs to enhance performance.

Limitations

Despite the promising results, our study has several limitations. First, the relatively small dataset may not fully capture the diversity of both AI-generated and human-written reviews, potentially limiting the generalizability of our findings. Second, our study focuses exclusively on Malayalam, restricting the applicability of the approach to other Dravidian and low-resource languages. Additionally, while some large language models, such as Gemma-2B, performed well, others underperformed, highlighting the need for further investigation into model selection and optimization for AI-generated text detection. A key challenge observed is the multilingual incapability of certain LLMs, which may stem from insufficient training data in Dravidian languages. Finally, the dataset may not encompass the full spectrum of AI-generated writing styles, which could affect the robustness of the classification models in real-world scenarios. To address these limitations, future work should explore more extensive and diverse datasets, use models with more parameters, extend the approach to other Dravidian languages, and refine model architectures to enhance performance in low-resource multilingual settings.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open-source models. Accessed: 2025-01-29.

Şule Öztürk Birim, Ipek Kazancoglu, Sachin Kumar Mangla, Aysun Kahraman, Satish Kumar, and Yigit Kazancoglu. 2022. Detecting fake reviews through topic modelling. *Journal of Business Research*, 149:884–900.

Gregorius Satia Budhi, Raymond Chiong, and Zuli Wang. 2021. Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. *Multimedia Tools and Applications*, 80:13079–13097.

Li-Chen Cheng, Yan Tsang Wu, Cheng-Ting Chao, and Jenq-Haur Wang. 2024. Detecting fake reviewers from the social context with a graph neural network method. *Decision Support Systems*, 179:114150.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Arkadipta De, Dibyanayan Bandyopadhyay, Baban Gain, and Asif Ekbal. 2021. A transformer-based approach to multilingual fake news detection in low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–20.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alessandro Gambetti and Qiwei Han. 2023. Combat ai with ai: Counteract machine-generated fake restaurant reviews on social media. *arXiv preprint arXiv:2302.07731*.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Jiwei Luo, Guofang Nan, Dahui Li, and Yong Tan. 2023. Ai-generated review detection. *Available at SSRN 4610727*.

Anirban Mukherjee. 2024. Safeguarding marketing research: The generation, identification, and mitigation of ai-fabricated disinformation. *arXiv preprint arXiv:2403.14706*.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Antonette Shibani, Faerie Mattins, Srivarshan Selvaraj, Ratnavel Rajalakshmi, and Gnana Bharathy. 2024. Tamil co-writer: Towards inclusive use of generative ai for writing support. In *LAK Workshops*, pages 240–248.

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

NAYEL@DravidianLangTech-2025: Character N-gram and Machine Learning Coordination for Fake News Detection in Dravidian Languages

Hamada Nayel[1,2], Mohammed Aldawsari[1], Hosahalli Lakshmaiah Shashirekha[3]

[1]Department of Computer Engineering and Information, Faculty of Engineering,
Prince Sattam Bin Abdulaziz University, Wadi Addawasir, Saudi Arabia
[2]Department of Computer Science, Faculty of Artificial Intelligence,
Benha University, Egypt
[3]Department of Computer Science, Mangalore University, India

Correspondence: hamada.ali@fci.bu.edu.eg

Abstract

This paper introduces the detailed description of the submitted model by the team NAYEL to Fake News Detection in Dravidian Languages shared task. The proposed model uses a simple character n-gram TF-IDF as a feature extraction approach integrated with an ensemble of various classical machine learning classification algorithms. While the simplicity of the proposed model structure, although it outperforms other complex structure models as the shared task results observed. The proposed model achieved a f1-score of 87.5% and secured the 5th rank.

1 Introduction

The growth of social media platforms have significantly contributed to the widespread issue of fake news across the globe. Fake news, which refers to intentionally deceptive or inaccurate information circulated through internet, presents serious challenges to public trust, governance, and societal health (Ashraf et al., 2022). While most research in fake news detection has concentrated on widely spoken languages like English, the distinct linguistic and cultural characteristics of regional languages have been largely overlooked (Nayel and Amer, 2021). Dravidian languages, spoken mainly in southern India, have received research attention in the context of automatically fake news detection (Subramanian et al., 2024; Devika et al., 2024; Subramanian et al., 2023).

Dravidian languages, such as Tamil, Telugu, Kannada, and Malayalam, are highly diverse in terms of their linguistic structures, including syntax, semantics, and morphology. These languages create unique difficulties for natural language processing (NLP) models, especially when applied to tasks like fake news detection. The variety in regional dialects, differences in scripts, and the influence of local culture further complicate the process of distinguishing between true and false information (Hegde et al., 2024, 2023).

This paper explores the submitted model to the Fake News Detection in Dravidian Languages shared task. This shared task is divided into two subtasks, the first subtask aimed at classifying a a given social media text into original or fake. While, the second subtask aimed at detecting the fake news from Malayalam News into five fake categories as well as original. Our team participated in the first subtask, and have submitted three runs.

The rest of the paper demonstrates the structure of the submitted model and the experimental results that have been produced in the devolvement phase.

2 Literature Review

Research works in fake news detection have gained interesting in last few years according to the massive usage of social media platforms.

Nayel and Amer (2021) used a simple Term Frequency-Inverse Document Frequency (TF-IDF) framework to extract the features of Urdu tweets and integrate with a linear classifier that achieved f1-score of 67.9% and outperformed all the submitted runs. The basic frame work that combines TF-IDF and ML algorithms has been used efficiently in various tasks such as text classification (Ashraf et al., 2024) and word level language identification (Ismail et al., 2022; Fetouh and Nayel, 2023). In the era of large language models (LLMs), they have been adapted for fake news detection (Hu et al., 2024; Su et al., 2024).

The research work has been done in fake news detection in Dravidian languages varies from the classical machine learning approaches (K et al., 2024), deep learning approaches (M et al., 2024) and transformer-based approach (Tabassum et al., 2024). K et al. (2024) explored character n-gram model with classical ML algorithms such as Linear Regression (LR), Support Vector Machines (SVMs), Naive Bayes (NB) and an ensemble model.

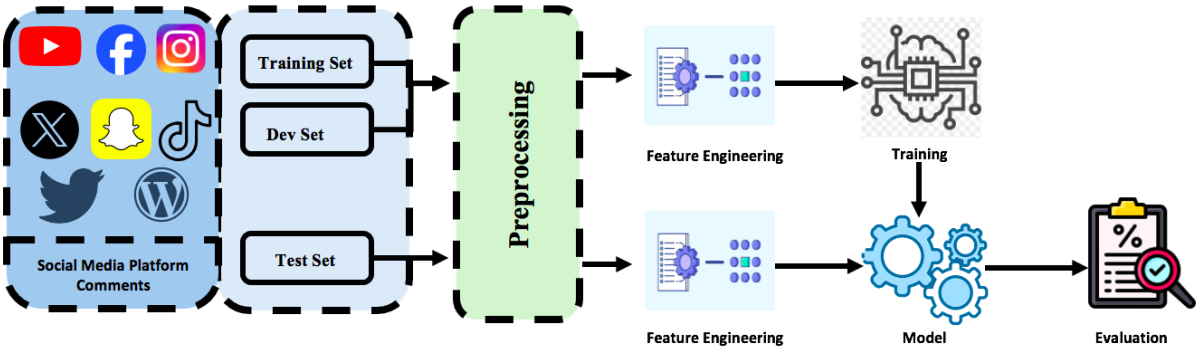


Figure 1: The structure of the proposed ML-based model

3 Dataset

The dataset has been used in this task was collected from various social media platforms such as X (formerly Twitter), Facebook etc. Detailed description of dataset, the methodology has been used to collect comments statistics and detailed analysis are given in (Subramanian et al., 2025). Table 1 shows the statistics of the dataset for both subtasks.

Task	Class	Train	Test	Dev
Task A	Fake	1599	507	406
	Original	1658	512	409
Task B	Half True	145	24	–
	False	1251	149	–
	Partly False	44	14	–
	Mostly False	242	63	–

Table 1: Statistics of the dataset

4 Methods

The proposed model composite of an ensemble of three base classifiers namely; SVMs, NB and linear classifier with Stochastic Gradient Descent (SGD) as an optimization algorithm. Majority voting mechanism has been used to combine the outputs of the base classifiers. As shown in figure 1, the general structure of the base classifier consists of dataset pre-processing, feature engineering, model training and evaluation.

- *Data pre-processing*
  Data pre-processing or text cleaning aims at omitting the unwanted texts such as stopwords, repeated letters, emojis and any uninformative tokens. In our model we employed a simple pre-processing procedure that removes the repeated characters, emojis and stopwords.

- *Feature Engineering*
  This phase involves feature extraction and selection. We employed a simple character n-gram TF-IDF for feature extraction. A wide range of n-grams including 3-gram, 4-gram, 5-gram and 6-gram have been extracted for the given text. Each token is consumed as a feature This approach reported improved results in .

- *Model Training*
  A set of classical machine learning classification algorithms have been implemented. The set of classifiers are: SVM, Naive Bayes and SGD. A voting-based ensemble model has been implemented using aforementioned classifiers as base classifiers. Ensemble learning is a machine learning technique that aggregates several individual models to produce more accurate predictions than a single model alone (Nayel and Shashirekha, 2017).

- *Evaluation*
  Performance evaluation of the model has been measured by f1-score, which is widely used in such cases of text classification. F1-score is a harmonic mean of precision (P) and recall (R) and is calculated as follows:

$$f1 - score = \frac{2PR}{R + P}$$

The general structure of voting-based ensemble model is shown in figure 2. The output of the base classifiers are input to a majority voting function.

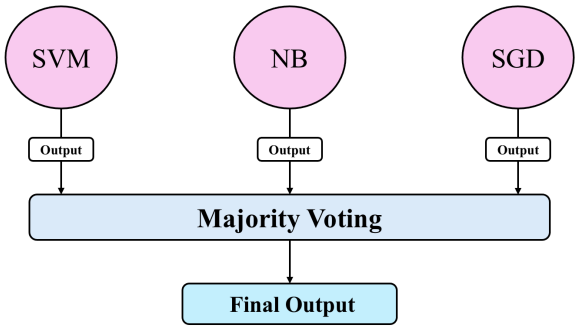


Figure 2: The structure of the majority voting ensemble model

5 Experimental Setting and Results

In this section experimental setting hyper parameters and running environment have been discussed. The code is freely available on GitHub repository[1].

A free package for `python` implementation of classical machine learning algorithms text features extraction `sklearn`[2] has been used to implement proposed model. In development phase and to fit the hyper parameters, the development set provided by shared task organizers has been used. To get the same results at each run, we utilized the package `random` and set the parameter `random_state` at 42.

The results of the applying the proposed models on development set are given in Table 2. The

Classifier	Precision	Recall	F1-score
SGD	0.89	0.86	0.87
NB	0.84	**0.88**	0.85
SVM	**0.90**	0.85	**0.88**
voting	0.89	0.86	**0.88**

Table 2: Results of the proposed model on development set

results on development set show that SGD-based model reported the minimum results in terms of all metrics. While, SVM outperforms other models in terms of precision. NB-based model reported the highest recall. SVM and voting classifiers reported the highest f1-score. Voting-based model, as it is clear, utilize the strength of all classifiers.

For test set, the majority voting based model outperforms all baseline model and reported f1-score of 0.875.

[1] https://github.com/hamadanayel/NAYEL_DRAVIDIAN

[2] https://scikit-learn.org

6 Conclusion

Fake news detection is a vital task in the era of social media expansion especially for low resources languages such as Dravidian languages. This work proposed a basic model that uses character n-gram TF-IDF and ML algorithms. The results obtained by the proposed model is a promising according to its simplicity and the low computational resources have been spent.

The model can be improved by applying more pre-processing steps as well as surveying more classical ML algorithms. In addition, LLMs can be tested in this case.

7 Limitations

The proposed model as clear uses character n-gram TF-IDF as a feature extraction approach, which is not efficient in the sense of semantic features. This is a lexicographical feature, while the meaning of the token is not used. In addition, The quality of dataset is a very important factor to develop an amenable systems.

LLMs can be applied to the task, but the limited computational resources access leads to applying the classical computational models such as ML and deep learning.

References

Nsrin Ashraf, Hamada Nayel, Mohammed Aldawsari, Hosahalli Shashirekha, and Tarek Elshishtawy. 2024. BFCI at AraFinNLP2024: Support vector machines for Arabic financial text classification. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 446–449, Bangkok, Thailand. Association for Computational Linguistics.

Nsrin Ashraf, Hamada Nayel, and Mohamed Taha. 2022. Misinformation detection in arabic tweets: A case study about covid-19 vaccination. *Benha Journal of Applied Sciences*, 7(5):265–268.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Ahmed M. Fetouh and Hamada Nayel. 2023. BFCAI at coli-tunglish@fire 2023: Machine learning based model for word-level language identification in code-mixed tulu texts. In *Working Notes of FIRE 2023 -*

Forum for Information Retrieval Evaluation (FIRE-WN 2023), Goa, India, December 15-18, 2023, volume 3681 of *CEUR Workshop Proceedings*, pages 205–212. CEUR-WS.org.

Asha Hegde, F Balouchzahi, Sharal Coelho, Shashirekha H L, Hamada A Nayel, and Sabur Butt. 2024. Coli@fire2023: Findings of word-level language identification in code-mixed tulu text. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 25–26, New York, NY, USA. Association for Computing Machinery.

Asha Hegde, Fazlourrahman Balouchzahi, Sharal Coelho, H. L. Shashirekha, Hamada A. Nayel, and Sabur Butt. 2023. Overview of coli-tunglish: Word-level language identification in code-mixed tulu text at FIRE 2023. In *Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation (FIRE-WN 2023), Goa, India, December 15-18, 2023*, volume 3681 of *CEUR Workshop Proceedings*, pages 179–190. CEUR-WS.org.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113.

Shimaa Ismail, Mai K. Gallab, and Hamada Nayel. 2022. BoNC: Bag of n-characters model for word level language identification. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 34–37, IIIT Delhi, New Delhi, India. Association for Computational Linguistics.

Manavi K, Sonali K, Gauthamraj K, Kavya G, Asha Hegde, and Hosahalli Shashirekha. 2024. MUCS@DravidianLangTech-2024: Role of learning approaches in strengthening hate-alert systems for code-mixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 252–256, St. Julian's, Malta. Association for Computational Linguistics.

Madhumitha M, Kunguma M, Tejashri J, and Jerin Mahibha C. 2024. TechWhiz@DravidianLangTech 2024: Fake news detection using deep learning models. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 200–204, St. Julian's, Malta. Association for Computational Linguistics.

Hamada Nayel and Ghada Amer. 2021. A simple n-gram model for urdu fake news detection. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021*, volume 3159 of *CEUR Workshop Proceedings*, pages 1150–1155. CEUR-WS.org.

Hamada Nayel and H. L. Shashirekha. 2017. Improving NER for clinical texts by ensemble approach using segment representations. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 197–204, Kolkata, India. NLP Association of India.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting fake news detection to the era of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1473–1490, Mexico City, Mexico. Association for Computational Linguistics.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_Punctuators@DravidianLangTech-EACL2024: Transformer-based approach for detection and classification of fake news in Malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186, St. Julian's, Malta. Association for Computational Linguistics.

AnalysisArchitects@DravidianLangTech 2025: BERT Based Approach For Detecting AI Generated Product Reviews In Dravidian Languages

Abirami Jayaraman abirami2210382@ssn.edu.in
Aruna Devi Shanmugam aruna2210499@ssn.edu.in
Dharunika Sasikumar dharunika2210459@ssn.edu.in
Bharathi B bharathib@ssn.edu.in

Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, Chennai, Tamil Nadu

Abstract

The shared task on Detecting AI-generated Product Reviews in Dravidian Languages is aimed at addressing the growing concern of AI-generated product reviews, specifically in Malayalam and Tamil. As AI tools become more advanced, the ability to distinguish between human-written and AI-generated content has become increasingly crucial, especially in the domain of online reviews where authenticity is essential for consumer decision-making. In our approach, we used the ALBERT, IndicBERT, and Support Vector Machine (SVM) models to classify the reviews. The results of our experiments demonstrate the effectiveness of our methods in detecting AI-generated content.

1 Introduction

The proliferation of AI-generated content has raised significant concerns across various domains, particularly in online reviews where authenticity is paramount for consumer decision-making. The Shared Task on Detecting AI-generated Product Reviews in Dravidian Languages(Premjith et al., 2025) addresses this issue by focusing on the detection of AI-generated reviews in Malayalam and Tamil. As AI tools become more sophisticated, distinguishing between human-written and AI-generated content has become increasingly challenging and crucial.

Online reviews play a critical role in influencing consumer behavior and purchasing decisions. However, the rise of generative AI has led to an increase in fake reviews, which can undermine consumer trust and distort market dynamics. Luo et al.(Luo et al., 2023) highlight the impact of AI-generated fake reviews on e-commerce platforms and propose a supervised learning approach to detect such reviews. Their study emphasizes the importance of developing robust detection methods to maintain the integrity of online reviews.

In this shared task, datasets containing both human-written and AI-generated reviews in Malayalam and Tamil were provided. The objective was to develop models capable of accurately classifying these reviews while addressing the specific challenges posed by Dravidian languages. Models including ALBERT, IndicBERT, and Support Vector Machine (SVM) classifiers were utilized in this approach.

The effectiveness of these models in detecting AI-generated reviews in Malayalam and Tamil was demonstrated through our experiments. Section 2 of this paper provides a brief summary about various works done in this filed. Section 3 provides the description of the datasets. Section 4 explains the methodology used. Section 5 provides detailed information about the implementation of the methodology. Section 6 provides a consolidated view of the results obtained in the test. Section 7 elaborates the shortcomings of these models. Section 8 concludes the paper.

2 Related Works

The study by Gupta and Jindal (Gupta et al., 2024) highlights the challenge of AI-generated fake reviews, which are produced using generative AI tools, complicating the integrity of online feedback systems. To combat this issue, various detection techniques are employed, including rule-based approaches that utilize predefined characteristics of fake reviews, graph-based techniques that analyze user-review relationships for anomalies, machine learning algorithms trained on review features to classify authenticity, and deep learning models that capture complex patterns in the data. The research emphasizes ongoing efforts to enhance detection accuracy and identifies key challenges, such as the

evolving tactics of those generating fake reviews, particularly through sophisticated AI-generated content.

This study by Jean Michel Sahut (Sahut et al., 2024) presents a novel supervised learning approach aimed at distinguishing between human-written reviews and those generated by AI. The study constructs various variables and employs an outlier detection method based on cumulative probability density to enhance detection accuracy. It demonstrates that the proposed method outperforms existing baseline techniques in identifying AI-generated reviews.

The study by Mudasir Ahmad Wani et al (Wani et al., 2024) introduces a framework utilizing deep learning algorithms and natural language processing (NLP) techniques to detect AI-generated spam reviews. The framework integrates multiple deep learning architectures, such as CNNs and LSTMs, and applies advanced NLP methods for thorough textual analysis, proving effective across diverse datasets.

The study by Jiwei et al (Mohammed and Ahmed, 2023) examines the impact of AI-generated reviews on consumer perceptions in online shopping. It discusses how AI tools can manipulate buyer behavior by producing reviews that may misrepresent products. While these reviews can enhance the shopping experience by summarizing key features, they also raise concerns about reliability and trust in e-commerce. The study emphasizes the need for vigilance regarding the integration of AI-generated content in online platforms, as it significantly influences market dynamics and consumer trust.

The paper by Anna Shcherbiak et al(Shcherbiak et al., 2024) investigates the effectiveness of various classifiers in distinguishing between texts generated by AI and those written by humans. The study employs a dataset comprising both AI-generated and human-written texts, utilizing advanced machine learning models to perform the classification task.

The research by Lorenz Mindner et al(Mindner et al., 2023) explores various features to detect AI-generated texts, including those rephrased by AI. The study uses a new text corpus and achieves high F1-scores in classifying both basic and advanced human-generated and AI-generated texts

Labels	Malayalam	Tamil
AI	406	401
HUMAN	394	408

Table 1: Split up of training data into 2 classes for Malayalam an Tamil

Labels	Malayalam	Tamil
AI	100	48
HUMAN	100	52

Table 2: Split up of testing data into 2 classes for Malayalam an Tamil

3 Dataset Description

The dataset used in this study consists of Tamil and Malayalam reviews, which include both human-written and AI-generated text. It is divided into training and validation sets to help with effective classification.

For the Malayalam dataset, there are 801 training samples, categorized into two labels: Human-written and AI-generated and are tabulated in Table 1. The test set contains 201 samples, which are used to evaluate the performance of the model. A detailed breakdown of this dataset can be seen in Table 2.

Similarly, the Tamil dataset has 809 training samples, also classified under the same two labels and are tabulated in Table 1. The test set consists of 101 samples, which help assess the accuracy of classification models. The distribution of this dataset is provided in Table 2.

4 Proposed Work

The goal of this paper is to classify Tamil and Malayalam text data to determine whether the content is human or AI-generated. Advanced transformer-based language models are employed for this task due to their efficiency and strong performance. The dataset, consisting of Tamil and Malayalam text samples with labels indicating "HUMAN" or "AI," is first preprocessed. Labels are encoded numerically, with "HUMAN" mapped

to 0 and "AI" to 1.

In this study, the three models, ALBERT(Lan et al., 2020), IndicBERT(Kakwani et al., 2020), and SVM, were used. After training, the model was evaluated on the test data to generate predictions. Metrics such as accuracy and macro F1-score were used to assess the model's ability to distinguish between human and AI-generated text.

5 Experimental Results

Implementation involves using ALBERT (A Light Bidirectional Encoder Representations from Transformers), Indic-BERT, and SVM (Support Vector Machine) models to classify Tamil and Malayalam text data as human-generated or AI-generated. The data preprocessing steps are largely similar for all three models, with differences arising primarily during the training process.

Labels are converted into numeric values using the label encoding. Specifically, the label "HUMAN" is encoded as 0, and "AI" is encoded as 1. This encoding step ensures that the data are compatible with machine learning models, which require numeric labels for classification tasks. After label encoding, the data is split into training and validation sets. Typically, 75% of the data is allocated for training, and the remaining 25% is reserved for validation. Text data is then tokenized using appropriate tokenizers for each model. Tokenization is the process of converting the text into smaller units (tokens), padding or truncating the sequences to a fixed length, and converting them into numerical representations. This ensures that the text data is in a format suitable for the respective models.

ALBERT model is a transformer-based architecture designed for efficient text classification tasks. The model and its tokenizer are loaded using the Hugging Face Transformers library. The tokenizer preprocesses the text data by converting it into a format compatible with ALBERT. This involves tokenizing the text, padding or truncating sequences to a fixed length, and converting the tokens into numerical values. The training process is configured using the Training Arguments class, where hyperparameters such as learning rate, batch size, number of epochs, and weight decay are specified.

Model used	Accuracy	Macro F-1 score
ALBERT	0.9136	0.9122
IndicBERT	0.9653	0.9651
SVM	0.8465	0.8465

Table 3: Performance on Tamil training dataset

Model used	Accuracy	Macro F-1 score
ALBERT	0.950	0.9499
IndicBERT	0.965	0.9649
SVM	0.775	0.7748

Table 4: Performance on Malayalam training dataset

Model is trained using the Trainer class, which handles the training loop, including backpropagation and optimization. Once the model is trained, it is evaluated on the validation set. The test data is tokenized similarly to the training data and passed through the trained ALBERT model to obtain predictions. These predictions are mapped back to their original labels ("HUMAN" or "AI") for interpretability.

The Indic-BERT model is another transformer-based model, specifically designed for Indic languages like Tamil and Malayalam. The training process for Indic-BERT follows a similar approach to ALBERT. The model and tokenizer are loaded from the Hugging Face library, and the text data is tokenized into a format suitable for the model. The training configuration is set up using the Training Arguments class, and the model is trained using the Trainer class, similar to the ALBERT approach. The evaluation process is the same as with ALBERT, where the model is tested on the validation set, and the predictions are mapped back to their original labels.

For the SVM model, the process differs from the transformer-based models. Instead of tokenizing the text data into tokens and using pre-trained models, SVM requires feature extraction. The text data is converted into numerical features using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency). These features are then used to train the SVM classifier.

The SVM model is trained with linear kernel. The training process involves finding the optimal hyperplane that separates the human and

Model used	Accuracy	Macro F-1 score
ALBERT	0.46	0.4375
IndicBERT	0.62	0.62
SVM	0.66	0.6594

Table 5: Performance on Tamil testing dataset

Model used	Accuracy	Macro F-1 score
ALBERT	0.885	0.8849
IndicBERT	0.5	0.333
SVM	0.68	0.6797

Table 6: Performance on Malayalam testing dataset

AI-generated text. The performance of the SVM model is evaluated using metrics such as accuracy and macro F1-score. Once trained, the SVM model is used to make predictions on the validation set. The predictions are then mapped back to the original labels ("HUMAN" or "AI").

The evaluation metrics used include accuracy and macro F1-score. Accuracy measures the overall correctness of the predictions, while the macro F1-score provides a balanced evaluation by considering both precision and recall for each class and averaging them. The performance of these models on Tamil training dataset is tabuated in Table 3 and on Malayalam training set is tabulated in Table 4

For all models, the results are saved to a TSV file, which includes the ID of each test instance and its corresponding predicted label. These results are then analyzed to assess the performance of each model in distinguishing between human and AI-generated text. Implementation of all these models are available in Github.[1]

6 Result

The performance of all the models on the Tamil testing dataset has been listed in Table 5, and on the Malayalam testing dataset in Table 6. From these tabulations, it can be inferred that better performance was achieved by SVM on the Tamil dataset and it's confusion matrix is available in Figure 1, while ALBERT performed better on the Malayalam dataset and it's confusion matrix is available in Figure 2. Although IndicBERT

[1]Github page : https://github.com/S-ArunaDevi06/AI_generated_review_detection/

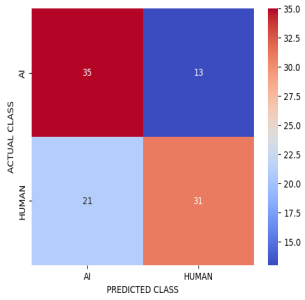


Figure 1: Confusion matrix for performance of SVM on Tamil testing dataset

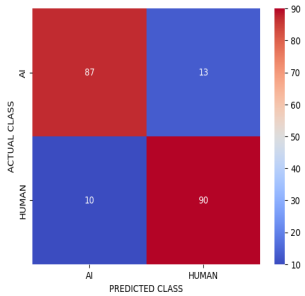


Figure 2: Confusion matrix for performance of ALBERT on Malayalam testing dataset

performed well on the training dataset for both languages, its performance on the testing dataset was comparatively lower for both languages.

By analysing the performance of SVM on tamil testing dataset, AI-written sentences that are misclassified as human-written sentences often contain structured, well-articulated language. By analysing the performance of ALBERT on Malayalam testing dataset, it can be observed that the model often struggles with sentences that have an informal or conversational tone. This model also struggles with sentences have transliterated words.

7 Limitations

In the ALBERT model, the maximum token length (512) can be restrictive for longer AI-generated content. If the dataset contains AI text from only one model (e.g., ChatGPT, GPT-3), ALBERT may overfit and fail on texts from newer AI models (e.g., Gemini, Claude).

Like ALBERT, IndicBERT might struggle to detect AI-generated text from newer AI models. SVM treats text as vectors without sequential information, it struggles with coherence and fluency patterns that distinguish AI from human

text. SVM works well for clearly separable classes. But if AI-generated and human text are very similar, SVM fails.

8 Conclusions

In this shared task, various BERT models as well as SVM were evaluated for classification tasks on Dravidian languages. The 23rd place was secured using SVM for Tamil, and the 9th place was secured using ALBERT for Malayalam in this shared task.

References

Richa Gupta, Vinita Jindal, and Indu Kashyap. 2024. Recent state-of-the-art of fake review detection: a comprehensive review. *The Knowledge Engineering Review*, 39:e8.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *Preprint*, arXiv:1909.11942.

Jiwei Luo, Guofang Nan, Dahui Li, and Yong Tan. 2023. Ai-generated review detection. *Available at SSRN 4610727*.

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. *Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT*, page 152–170. Springer Nature Singapore.

Ennaouri Mohammed and Zellou Ahmed. 2023. Machine learning approaches for fake reviews detection: A systematic literature review. *Journal of Web Engineering*.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Jean Michel Sahut, Michel Laroche, and Eric Braune. 2024. Antecedents and consequences of fake reviews in a marketing approach: An overview and synthesis. *Journal of Business Research*, 175:114572.

Anna Shcherbiak, Hooman Habibnia, Robert Böhm, and Susann Fiedler. 2024. Evaluating science: A comparison of human and ai reviewers. *Judgment and Decision Making*, 19:e21.

Mudasir Ahmad Wani, Mohammed ElAffendi, and Kashish Ara Shakil. 2024. Ai-generated spam review detection framework with deep learning algorithms and natural language processing. *Computers*, 13(10).

AnalysisArchitects@DravidianLangTech 2025: Machine Learning Approach to Political Multiclass Sentiment Analysis of Tamil

Abirami Jayaraman **Aruna Devi Shanmugam** **Dharunika Sasikumar**
abirami2210382@ssn.edu.in aruna2210499@ssn.edu.in dharunika2210459@ssn.edu.in
Bharathi B
bharathib@ssn.edu.in

Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam, Chennai, Tamil Nadu

Abstract

Sentiment analysis is recognized as an important area in Natural Language Processing (NLP) that aims at understanding and classifying opinions or emotions in text. In the political field, public sentiment is analyzed to gain insight into opinions, address issues, and shape better policies. Social media platforms like Twitter (now X) are widely used to express thoughts and have become a valuable source of real-time political discussions. In this paper, the shared task of Political Multiclass Sentiment Analysis of Tamil tweets is examined, where the objective is to classify tweets into specific sentiment categories. The proposed approach is explained, which involves preprocessing Tamil text, extracting useful features, and applying machine learning and deep learning models for classification. The effectiveness of the methods is demonstrated through experimental results and the challenges encountered while working on the analysis of Tamil political sentiment are discussed.

1 Introduction

Sentiment analysis has become a pivotal area within Natural Language Processing (NLP), focusing on identifying sentiments expressed in text. This capability is particularly valuable in political contexts, where understanding public sentiment can guide decision-making and policy improvements. Social media platforms, such as Twitter (now X), serve as rich sources of real-time data, offering diverse perspectives on political issues. Analyzing public discourse on these platforms provides essential information on the emotions and opinions of the population.

This study specifically focuses on the analysis of political multiclass sentiment in Tamil tweets. Unlike general sentiment analysis, this task requires the classification of tweets into multiple sentiment categories tailored to the political domain. The complexity of working with Tamil text is further compounded by challenges such as code-mixing, spelling variations, and a scarcity of annotated datasets. To address these challenges, various models, including machine learning and deep learning techniques, were tested. The Naïve Bayes algorithm was used for its simplicity and effectiveness in text classification, Support Vector Machines (SVM) were used for their robustness in handling high-dimensional data, and Long-Short-Term Memory (LSTM) networks were used to capture sequential patterns in text data.

Sentiment analysis was applied effectively to political contexts in various languages. For example, Gunhal's study (Gunhal, 2023a) on the sentiment surrounding Karnataka's elections used transformer-based models to analyze Twitter data, achieving significant precision in predicting electoral results. This is consistent with previous research demonstrating the predictive influence of Twitter sentiment on electoral results. Furthermore, previous studies by Krishnan(Krishnan et al.) have highlighted the importance of sentiment analysis in understanding public reactions to political events and policies.

The paper is structured as follows: Section 2 presents an overview of related works, summarizing existing research on sentiment analysis and Tamil text classification. Section 3 describes the dataset used, including its sources and preprocessing steps. In Section 4, we detail our methodology encompassing data preprocessing, feature extraction, and the models employed. Section 5 provides detailed description about the implementation of the work proposed. Section 6 presents our experimental findings and performance metrics used for evaluation. Section 7 elaborates the shortcomings of these models in general. Finally, Section 8 concludes with a summary of the work done and it's

result in Tamil political sentiment analysis.

2 Related Works

The study by Pranav Gunhal(Gunhal, 2023b) investigates sentiment classification surrounding the 2023 Karnataka elections using transformer-based models, particularly IndicBERT. This research explores innovative data collection and augmentation techniques to analyze Twitter sentiment, classifying posts as positive, negative, or neutral. The findings demonstrate the potential of these models in forecasting electoral outcomes and capturing sentiment trends in Indian politics, highlighting their value for political stakeholders in future elections.

Rajasekar et al(Rajasekar and Geetha, 2023) present a robust deep learning model specifically designed for analyzing Tamil tweets. The authors aim to improve sentiment classification accuracy using deep convolutional neural networks (CNNs). The model significantly outperforms traditional methods, demonstrating its effectiveness in capturing sentiment nuances within Tamil language tweets.

Chakravarthi et al. (Chakravarthi et al., 2025) presented a machine learning-based approach for political multiclass sentiment analysis of Tamil X (Twitter) comments as part of the DravidianLangTech shared task. Their study explores various machine learning techniques to classify sentiments, contributing to the development of sentiment analysis for underrepresented languages.

The research by Sajeetha et al (Thavareesan and Mahesan, 2019) explores various machine learning approaches for sentiment analysis in Tamil texts, including lexicon-based and supervised learning methods. By comparing these techniques, the study identifies effective strategies for feature representation and sentiment classification, highlighting the need for customized methodologies to analyze Tamil sentiments.

The work of Sharmista(Sharmista and Ramaswami, 2020) focuses on the extraction of opinions within Tamil language social media reviews, addressing the scarcity of research in this area. The study aims to improve the understanding of consumer sentiments among Tamil speakers, providing foundational insights that can be leveraged for broader sentiment analysis applications.

Sentiment	Training	Testing
Positive	578	75
Negative	407	46
Neutral	638	70
Opinionated	1316	172
Substantiated	412	51
Sarcastic	790	108
None of the above	172	25

Table 1: Split up of training and testing data into 7 sentiment classes

The research of Ponnusamy et al(Ponnusamy et al., 2023) tackles the complexities of detecting sentiments in code-mixed comments between Tamil and Tulu languages on social media platforms. It proposes preprocessing techniques to improve model performance and offers insights into handling linguistic diversity in sentiment analysis tasks. Hetu et al.(Bhavsar and Manglani, 2019) built and proposed a model in sentiment analysis on twitter data . They classify the emotions based on positive and negative reviews. This model gives high accuracy on large dataset.

3 Dataset Description

The training dataset consists of 4353 Tamil tweets. These are categorized into 7 classes: Positive, Negative, Neutral, Opinionated, Substantiated, Sarcastic, and None of the above. The distribution of the data across these sentiment classes is listed in Table 1. From Table 1, we can infer distribution of data into these classes is not equal.The test set consists 544 Tamil tweets.

4 Proposed Work

The objective of this work is to develop an advanced text classification system capable of categorizing Tamil text into predefined sentiment categories, such as opinionated, neutral, substantiated, positive, sarcastic, and negative. The task involves utilizing machine learning and deep learning techniques to effectively capture the nuances in Tamil text and accurately classify the content based on sentiment.

The first step involves preprocessing the data, which includes extracting text and sentiment labels

from the dataset. The sentiment labels are encoded into numerical values to facilitate the training process. Typically, the labels are transformed using techniques such as Label Encoding or One-Hot Encoding. The dataset is then divided into training and validation sets, with a common split of 80% for training and 20% for validation.

The text processing step includes tokenizing the text, where words are converted into sequences of integers, and then padding or truncating the sequences to a fixed length. This process ensures that all input sequences are of the same length, which is essential for feeding the data into machine learning models.

For model development, various machine learning and deep learning algorithms are explored. Traditional approaches like Naive Bayes and Support Vector Machines (SVM) and LSTM (Long-Short-Term Memory) models are used for this text classification task.

Training of the model involves optimizing hyperparameters such as learning rate, batch size, and the number of epochs to achieve the best performance. The model is then evaluated using metrics like accuracy and Macro F1-score to assess its ability to correctly classify the text into the respective sentiment categories.

Finally, once the model is trained and evaluated, it is used to make predictions on unseen data. The predicted sentiment labels are mapped back to their corresponding categories, and the results are saved in a structured format such as CSV or JSON for further analysis.

5 Experimental Results

The implementation of the proposed approach is carried out in multiple stages, including data preprocessing, feature extraction, model training, and prediction. Initially, the text in the dataset is cleaned and standardized through preprocessing. All text is converted to lowercase to maintain uniformity, and special characters, punctuation, and other non-alphanumeric symbols are removed using regular expressions to reduce noise. Sentiment labels are then encoded into numerical values to facilitate model training. Given the challenges associated with Tamil text, such as spelling variations and code-mixing, careful preprocessing is performed to improve classification accuracy.

For feature extraction, two different techniques are employed: Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW) using CountVectorizer. The TF-IDF representation is used to capture important textual features by considering word frequency while reducing the impact of common words, thereby enhancing the model's ability to distinguish between sentiment classes. The BoW model, on the other hand, represents text as a sparse matrix of token counts, providing a straightforward yet effective approach for text classification tasks.

To classify the preprocessed text, two machine learning models—Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB)—are experimented with. The SVM classifier, using a linear kernel, is trained on TF-IDF features, leveraging its ability to handle high-dimensional text data and provide robust decision boundaries. The MNB model, a probabilistic classifier well-suited for text classification, is trained using BoW features and efficiently processes word frequency-based data. The trained models are then used to predict sentiments on the test dataset, where the same preprocessing and feature extraction techniques are applied to ensure consistency.

Additionally, Long-Short-Term Memory (LSTM) networks are utilized for classifying Tamil text data into various sentiment categories, including opinionated, neutral, substantiated, positive, sarcastic, and negative. The dataset is typically stored in a CSV file and read into a pandas DataFrame for preprocessing. The sentiment labels, representing different categories, are numerically encoded using Label Encoding. This transformation ensures compatibility with machine learning models by converting categorical labels into numeric values.

Subsequently, the dataset is split into training and testing sets using the train-test-split function from scikit-learn. A typical split ratio of 80% for training and 20% for testing is used, ensuring that the model is trained on a large portion of the data while its performance is evaluated on unseen data.

Model	Macro F1-Score	Accuracy
SVM	0.29	0.3566
Naïve Bayes	0.31	0.3566
LSTM	0.2282	0.3056

Table 2: Performance of Training Data

Model	Macro F1-Score	Accuracy
SVM	0.2747	0.35
Naïve Bayes	0.2726	0.35
LSTM	0.2585	0.32

Table 3: Performance on Testing Data

The text data is then preprocessed using a Tokenizer, which converts the text into sequences of integers. These sequences are either padded or truncated to a fixed length, ensuring that all input sequences have the same length. This step is necessary for feeding the data into an LSTM model, which requires a consistent input shape. Padding ensures that shorter sequences are extended, while longer sequences are truncated to maintain uniformity.

The LSTM model is then defined using the Sequential API from Keras. The model consists of an Embedding layer, which converts the input sequences into dense vectors, followed by an LSTM layer that captures temporal dependencies in the text. Dense layers with ReLU activation are included for hidden layers, along with a final Dense layer containing a softmax activation function for multi-class classification. The softmax activation is used to output probabilities for each sentiment class, and the class with the highest probability is selected as the predicted sentiment.

Once the model is trained, it is evaluated using the test set. The test data is tokenized and padded in the same manner as the training data, and the model's predictions are generated. The predicted labels, initially in numeric form, are mapped back to their corresponding sentiment categories using a predefined label mapping. The predictions are then saved to a CSV file, along with the original text and predicted labels.

Finally, the predicted sentiment labels are decoded back into their respective categories and stored for evaluation. The performance of the models is assessed using metrics such as accuracy and macro F1-score, which provide insights into overall classification performance and are tabulated in Table 2. Implementation of these models are available in github.[1]

[1] https://github.com/Dharunika-07/Political_sentiment_analysis

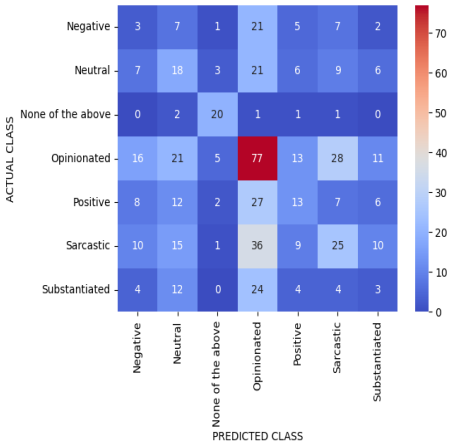


Figure 1: Confusion matrix for performance of SVM on testing dataset

6 Result

The performance of three machine learning models— Support Vector Machine (SVM), Naive Bayes (NB), and Long-Short-Term Memory (LSTM)— was evaluated on the Tamil Political Sentiment Analysis dataset. The results were presented in terms of macro F1-score and accuracy.

A slightly higher macro F1-score (0.2747) was achieved by the SVM model compared to Naive Bayes (0.2726) and LSTM (0.2585), while accuracies of 0.35, 0.35, and 0.32 were attained, respectively. The performance can be viewed in Table 3 and the confusion matrix for SVM is available in Figure 1. From the confusion matrix (Figure 1), we can say that the number correctly predicted sentences is higher for Opinionated class.

The model struggles to identify content that is neutral or factual and tends to assume it is opinion-based. This indicates that the model finds it challenging to distinguish between objective information and subjective language. Additionally, the Substantiated category was frequently confused with Neutral and Opinionated, suggesting that the model may have difficulty recognizing content that is supported by evidence.

7 Limitations

SVM is best suited for binary classification. Handling 7 classes requires one-vs-one or one-vs-all, which increases computation and may cause class imbalances. It also struggles with non-linear text relationships. SVM struggles if emotions in text are non-linearly separable. Emotions like sarcasm require deeper contextual understanding, which SVM lacks.

Naïve Bayes assumes that words occur independently, which is not true for Tamil. Tamil sentences have agglutination (words are formed by joining morphemes), affecting the assumption. It also struggles with classification between sarcasm neutrality.

Training LSTMs for Tamil takes time and GPU power. Tamil has long, context-dependent sentences, which may not be fully captured by a simple LSTM. It also needs large labeled data and requires expensive training.

8 Conclusions

In this task, we have secured the 15^{th} position. In this paper, a machine learning approach for Political Multiclass Sentiment Analysis of Tamil tweets is presented. SVM, Naïve Bayes, and LSTM models were applied, all demonstrating similar accuracy, with SVM slightly outperforming in terms of macro F1-score. Despite challenges such as code-mixing and limited annotated data, Tamil political sentiments were effectively classified using these methods. Future work could involve the exploration of deep learning techniques and data augmentation to address class imbalance and improve accuracy. This study contributes to sentiment analysis in Indian languages and provides insights for analyzing political discourse on social media.

References

Hetu Bhavsar and Richa Manglani. 2019. Sentiment analysis of twitter data using python. *International Research Journal of Engineering and Technology (IRJET)*, 6(3):510–511.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Pranav Gunhal. 2023a. Analyzing political sentiment of indic languages with transformers.

Pranav Gunhal. 2023b. Sentiment analysis in indian elections: unravelling public perception of the karnataka elections with transformers. *International Journal of Artificial Intelligence & Applications*, 14(5):41–55.

V Gokula Krishnan, Pinagadi Venkateswara Rao, J Deepa, and V Divya. Twitter sentiment analysis using ensemble classifiers on tamil and malayalam languages.

Kishore Kumar Ponnusamy, Charmathi Rajkumar, Prasanna Kumar Kumaresan, Elizabeth Sherly, and Ruba Priyadharshini. 2023. Vel@ dravidianlangtech: Sentiment analysis of tamil and tulu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 211–216.

M Rajasekar and Angelina Geetha. 2023. Sentiment analysis of tamil tweets using deep convolution neural networks. In *2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI)*, pages 1–5. IEEE.

A Sharmista and Dr M Ramaswami. 2020. Sentiment analysis on tamil reviews as products in social media using machine learning techniques: A novel study. *Madurai Kamaraj University Madurai-625*, 21.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on industrial and information systems (ICIIS)*, pages 320–325. IEEE.

TEAM_STRIKERS@DravidianLangTech2025: Misogyny Meme Detection in Tamil Using Multimodal Deep Learning

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2], **Mohamed Arsath H**[1], **Ramya K**[1], **Ragav R**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{mohamedarsathh.22aid, ramyak.22aid}@kongu.edu
ragavr.22aid@kongu.edu

Abstract

This study focuses on detecting misogynistic content in memes under the title Misogynistic Meme Detection Using Multimodal Deep Learning. Through an analysis of both textual and visual components of memes, specifically in Tamil, the study seeks to detect misogynistic rhetoric directed towards women. Preprocessing and vectorizing text data using methods like TF-IDF, GloVe, Word2Vec, and transformer-based embeddings like BERT are all part of the textual analysis process. Deep learning models like ResNet and EfficientNet are used to extract significant image attributes for the visual component. To improve classification performance, these characteristics are then combined in a multimodal framework employing hybrid architectures such as CNN-LSTM, GRU-EfficientNet, and ResNet-BERT. The classification of memes as misogynistic or non-misogynistic is done using sophisticated machine learning and deep learning approaches. Model performance is evaluated using metrics like as Accuracy, Precision, Recall, F1-Score, and Macro Average F1-Score. This study shows how multimodal deep learning can effectively detect and counteract negative narratives about women in digital media by combining natural language processing with image classification.

1 Introduction

The causes that propel the terrible part of misogyny online are inciting animosity and discriminating ideas toward women. Online communication is dominated by memes, which are extremely hard to spot because of the way they use both overt and covert misogynistic rhetoric in their words and visuals. Therefore, it is necessary to spot misogynistic material in memes. With an emphasis on examining both linguistic and visual elements, this study aims to categorize Tamil memes as either misogynistic or non-misogynistic. Misogynistic terms and phrases hidden in meme material will be detected using sophisticated natural language processing algorithms. Textual input is converted into intelligible representations for classification using transformer-based models like BERT and vectorization techniques like TF-IDF, GloVe, and word2vec. Relevant image features will be extracted for the visual analysis using deep learning models such as ResNet and EfficientNet. Modern multimodal deep learning models are integrated in the study, and metrics like Accuracy, Precision, Recall, F1-Score, and Macro F1-Score are used to assess how well they perform. This exacting approach guarantees a thorough comprehension of linguistic subtleties and cultural settings. The study advances AI strategies for addressing misogyny in regional languages by concentrating on Tamil. The groundwork for future initiatives to combat gender-based discrimination in online spaces is laid by this work, which also emphasizes the value of automated technologies in promoting digital civility.

2 Literature Survey

Cuervo and Parde (2022) highlights the paucity of research on multimodal systems intended to identify misogynistic content. Misogynistic memes are a common problem on social media that combine graphics with disparaging text to spread damaging messages. The authors apply contrastive learning in the context of SemEval 2022 Task 5, which is concerned with identifying sexist memes, by utilizing OpenAI's CLIP model, which is well-known for its efficacy in multimodal tasks. Even if the built model doesn't perform at its best, the tests offer insightful exploratory information that advances our knowledge of how to identify misogynistic content in memes.

Rizzi et al. (2023) explore methods for identifying misogynistic content in social media memes. The study highlights the introduction of a bias es-

timation technique and a Bayesian optimization strategy, resulting in a 61.43% improvement in prediction accuracy. They assess multiple unimodal and multimodal approaches, addressing challenges associated with specific meme archetypes. The authors emphasize the necessity for further research to mitigate model biases and enhance detection techniques.

Multimodal hate content detection has become more difficult due to the development of misogynistic memes. Multimodal misogyny is more difficult to detect than text-based sexism, even when using balanced datasets such as MAMI, which has 12,000 annotated memes Singh et al. (2023). Even with improvements, contextual ambiguity remains a challenge for models. However, performance is greatly enhanced by domain-specific pretraining, especially when BERT is used in conjunction with attention-based techniques.

In today's digital age, memes have emerged as a popular medium for online expression, humor, sarcasm, and social commentary. Yet, beneath their surface, they often carry troubling elements like misogyny, gender-based bias, and harmful stereotypes. To address these issues, Ponnusamy et al. (2024) delves into the world of online misogyny among Tamil and Malayalam-speaking communities, creating an annotated dataset with comprehensive guidelines. By analyzing memes, the authors reveal the complexities of gender bias and stereotypes, highlighting their manifestations and impact. This dataset and its detailed guidelines are invaluable resources for understanding the prevalence, origins, and nuances of misogyny, aiding researchers, policymakers, and organizations in formulating effective strategies to combat gender-based discrimination and promote equality and inclusivity. It provides profound insights that inform strategies for fostering a more equitable and safe online environment. This study is a crucial step in raising awareness and tackling gender-based discrimination in the digital realm.

A thorough summary of the shared work at LT-EDI@EACL 2024, which focused on classifying troll memes and misogynistic content in Tamil and Malayalam, may be found at Chakravarthi et al. (2024). According to the study, 52 teams entered the competition, and four systems—three for Malayalam and four for Tamil meme classification—were presented. The results of the shared work highlight the prevalence of troll and misogynistic content on the internet today and investigate the computational methods used to identify it. For Tamil and Malayalam, the best-performing model received macro F1 scores of 0.73 and 0.87, respectively.

Online sexism is a serious social problem that turns digital spaces into unfriendly places for women. In order to overcome this difficulty, Hashmi et al. (2024) suggest a strategy for identifying misogynous content in bilingual (English and Italian) online conversations. Explainable artificial intelligence (LIME), multilingual fine-tuned transformers, and FastText word embeddings are all used in the study. Their strategy performs better on important measures like accuracy and F1-score, demonstrating the possibility of cutting-edge approaches to successfully tackle online misogyny.

Angeline et al. (2022) combine BERT embeddings with Long Short-Term Memory (LSTM) networks to present a novel method for identifying sexist discourse on social media. While LSTM achieves an accuracy of 86.15% in capturing long-term dependencies in text, their study focuses on the contextual interpretation of tweets using BERT. This demonstrates how well deep learning models detect hazardous information on Twitter and other networks.

Habash et al. (2022) created a deep learning system to identify misogynistic memes. VisualBERT and two MMBT models are among the ensemble of multi-modal models used by the system. The two subtasks it tackles are identifying sexist memes (sub-task A) and classifying them into four categories: violence, objectification, shaming, and stereotyping (sub-task B). In sub-task A, their system outperformed the baseline model by a wide margin with an F1-score of 0.722.

Mahadevan et al. (2022) suggested a feature extraction-based method for detecting misogynous memes utilizing transformer models such as BERT and RoBERTa. By combining textual and visual components from memes, their multimodal training approach improves misogyny detection. The system demonstrated its efficacy in practical applications by performing remarkably well, placing fourth in Subtask A and ninth in Subtask B.

Pro-Cap, a novel probing-based captioning technique for hostile meme detection, was presented by Cao et al. (2023). It uses pre-trained vision-language models (PVLMs) without any fine-tuning. Pro-Cap generates image captions that contain crucial information for identifying hateful content by leveraging queries linked to hateful content to pro-

voke a frozen PVLM. This method's effectiveness and generalizability were validated by its good performance on three benchmarks.

3 Task Description

The shared task focuses on detecting misogynistic content in Tamil-language memes. Our objective is to categorize a dataset of misogynistic and non-misogynistic memes into two groups: Misogynistic and Non-Misogynistic. In order to develop models that can precisely detect misogynistic content while addressing issues specific to Tamil memes, this work entails assessing both textual and visual features. Chakravarthi et al. (2025) Our system showcased its performance in this demanding multimodal classification test by securing 17th place out of 118 participants.

4 Dataset Description

The dataset contains a total of 1,776 memes, split into three subsets 1,136 for training, 356 for testing, and 284 for development make up the dataset employed in this study. Three essential elements are present in every meme in the dataset: an image ID, a label, and transcriptions. The label denotes whether the meme is categorized as Misogynistic or Non-Misogynistic the picture ID is the unique identifier for each meme image, and the transcriptions are the textual material included in the meme. In order to aid in the creation and assessment of machine learning models intended to identify harmful online content, this dataset was created especially for the categorization of misogynistic content in memes, with an emphasis on Tamil language memes.

Dataset	No. of Memes
Train	1136
Validation	284
Test	356

Table 1: Tamil Dataset Description

5 Data Pre-processing

The data preprocessing pipeline for Misogynistic Meme Detection in Tamil using Multimodal Deep Learning is designed to efficiently handle both textual and visual data: text preprocessing involves converting Tamil text to lowercase, then removing URLs, HTML tags, special characters, and

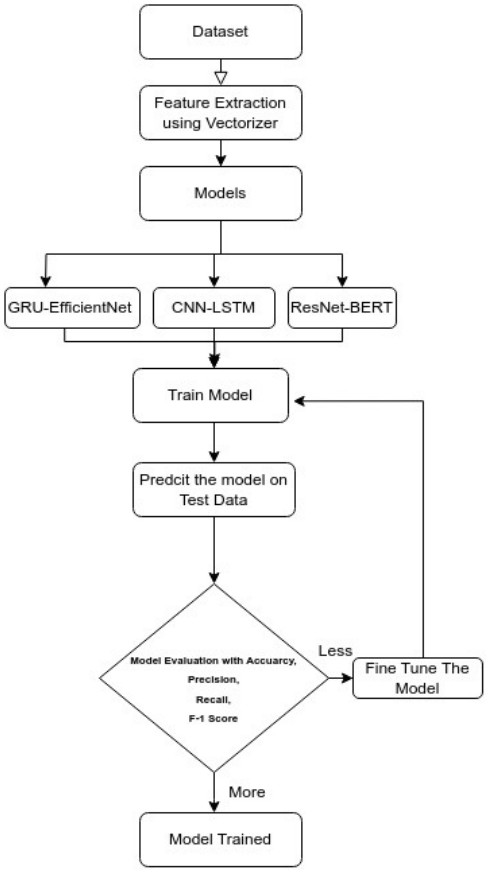


Figure 1: Proposed System Workflow

numbers to preserve meaningful content; tokenization is carried out using a Tamil-specific tokenizer, and stopword removal is applied using a curated list of Tamil stopwords; Tamil is a morphologically rich language, so stemming and lemmatization are handled using TamilMorph to normalize words; short words and duplicate words within a sentence are eliminated to reduce noise; image preprocessing involves resizing meme images to a fixed size, normalizing pixel values, and applying data augmentation techniques like rotation, flipping, and brightness adjustments to improve generalization.Tesseract OCR is used to extract Tamil text embedded in images using OCR preprocessing, and the retrieved text is then cleaned using the same procedure. Lastly, multimodal preprocessing improves the accuracy of sexist content detection in Tamil memes by combining linguistic and visual data to guarantee high-quality input for deep learning models.

6 Model Evaluation

The objective of this project is to classify memes as Misogynistic or Non-Misogynistic using a multimodal approach, incorporating both textual and

visual components. The trained model combines Natural Language Processing (NLP) techniques with computer vision methods to process and classify memes.

This experiment used a multimodal deep learning system that combined computer vision and NLP techniques to classify memes as misogynistic or non-misogynistic. CNNs like ResNet and EfficientNet extracted image features, while LSTM and GRU models processed Tamil-English code-mixed text with vectorization techniques like TF-IDF, GloVe, and Word2Vec. The CNN-LSTM model achieved the highest accuracy of 77.1%, leveraging CNN's spatial feature extraction and LSTM's ability to capture long-range text dependencies. The GRU-EfficientNet model followed with 76.8% accuracy, while the ResNet-BERT model scored 72.9%, likely due to BERT's limitations with code-mixed text. Models relying solely on text or image features failed to capture subtle contextual aspects, especially in cases involving sarcasm or culturally specific expressions. Despite these challenges, multimodal fusion effectively captured both textual and visual cues, proving to be a powerful strategy for detecting harmful online content, as demonstrated by accuracy, precision, recall, and macro average F1-score.

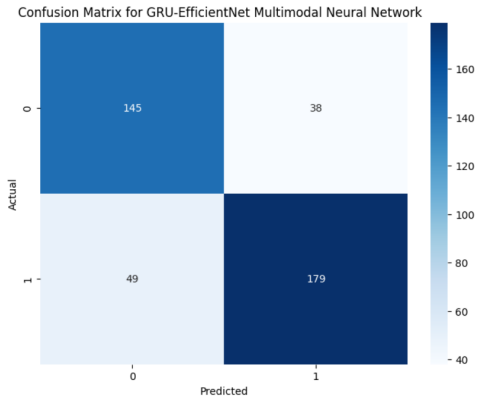


Figure 2: Confusion Matrix for GRU-EfficientNet

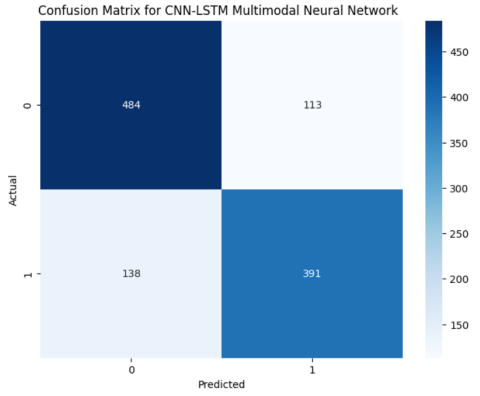


Figure 3: Confusion Matrix for CNN-LSTM

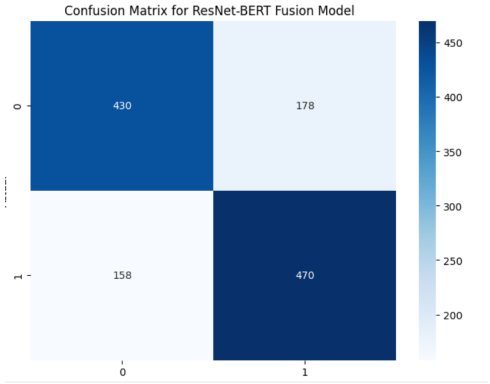


Figure 4: Confusion Matrix for ResNet-BERT

Model	Accuracy	F1 Score
GRU-EfficientNet Multimodal Neural Network	76.8%	0.62
CNN-LSTM Multimodal Neural Network	77.1%	0.68
ResNet-BERT Fusion Model	72.9%	0.63

Table 2: Model Accuracy and Macro Average F1 Score

7 Limitations

Despite encouraging outcomes, the models had trouble with complex situations where feature fusion was occasionally insufficient, such as irony, subtle misogyny, and culturally distinctive emotions. The reduced accuracy of the ResNet-BERT model was probably caused by BERT's shortcomings when it came to code-mixed Tamil-English text that lacked domain-specific fine-tuning. Furthermore, the models' reliance on huge datasets and powerful computers may restrict their scalability. By using contrastive learning strategies, bigger and more varied datasets, and refined transformer-based models specifically designed for Tamil social media material, future research can overcome these constraints.

8 Conclusion

This project combined picture classification and natural language processing (NLP) to assess both visual and linguistic features, resulting in a multimodal deep learning framework for misogynistic meme detection. The CNN-LSTM Multimodal Neural Network achieved the highest accuracy of 77.1% out of all the architectures we examined, followed by the GRU-EfficientNet model. Accuracy and macro F1-score demonstrated that models that combined text and image features performed better than single-modal approaches. These findings demonstrate the effectiveness of multimodal learning for challenging categorization tasks and show how sophisticated deep learning techniques can enhance the detection of dangerous content.

References

R. S . Angeline, D. Nurjanah, and H. Nurrahmi. 2022. Misogyny speech detection using long short-term memory and bert embeddings. In *2022 4th International Conference on Informatics and Computational Sciences (ICOIACT)*, pages 155–159.

R. Cao, M. S. Hee, A. Kuek, W. H. Chong, R. K.-W. Lee, and J. Jiang. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Cuervo and N. Parde. 2022. Exploring contrastive learning for multimodal detection of misogynistic memes. pages 785–792.

M. Habash, Y. Daqour, M. Abdullah, and M. Al-Ayyoub. 2022. Ymai at semeval-2022 task 5: Detecting misogyny in memes using visualbert and mmbt multimodal pre-trained models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 780–784.

E Hashmi, S. Y. Yayilgan, M. M. Yamin, and M. Ullah. 2024. Enhancing misogyny detection in bilingual texts using explainable ai and multilingual fine-tuned transformers. *Complex & Intelligent Systems*, 11(1).

S. Mahadevan, S. Benhur, R. Nayak, M. Subramanian, K. Shanmugavadivel, K. Sivanraju, and B. R. Chakravarthi. 2022. Transformers at semeval-2022 task 5: A feature extraction based approach for misogynous meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 550–554.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Rizzi, F. Gasparini, A. Saibene, P. Rosso, and E. Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing and Management*, 60:103474.

Singh, A. Haridasan, and R. J. Mooney. 2023. Female astronaut: Because sandwiches won't make themselves up there: Towards multimodal misogyny detection in memes.

KCRL@DravidianLangTech 2025: Multi-Pooling Feature Fusion with XLM-RoBERTa for Malayalam Fake News Detection and Classification

Fariha Haq, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, Golam Sarwar Md. Mursalin, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u1904051, u1904077, u1804128}@student.cuet.ac.bd,
sarwarmursalin1015@gmail.com, muhammad_ikhan@cuet.ac.bd

Abstract

The rapid spread of misinformation on social media platforms necessitates robust detection mechanisms, particularly for languages with limited computational resources. This paper presents our system for the DravidianLangTech 2025 shared task on Fake News Detection in Malayalam YouTube comments, addressing both binary and multiclass classification challenges. We propose a Multi-Pooling Feature Fusion (MPFF) architecture that leverages [CLS] + Mean + Max pooling strategy with transformer models. Our system demonstrates strong performance across both tasks, achieving a macro-averaged F1 score of 0.874, ranking 6th in binary classification, and 0.628, securing 1st position in multiclass classification. Experimental results show that our MPFF approach with XLM-RoBERTa significantly outperforms traditional machine learning and deep learning baselines, particularly excelling in the more challenging multiclass scenario. These findings highlight the effectiveness of our methodology in capturing nuanced linguistic features for fake news detection in Malayalam, contributing to the advancement of automated verification systems for Dravidian languages.

1 Introduction

Social media platforms have turned out to be primary channels of information dissemination, generating massive volumes of data that require sophisticated techniques of analysis for verification of content authenticity (Farsi et al., 2024). The term "fake news" encompasses content spread without verification, published uncritically, and deliberately disseminated to cause social disorder (Majumdar et al., 2021). With the proliferation of social media platforms, manually verifying the authenticity of each piece of information has become increasingly challenging (Yigezu et al., 2024). The increasing impact of fake news on public opinion has highlighted the need for advanced detection systems, especially for low-resource languages. This work addresses the challenge of fake news detection in Malayalam YouTube comments through two distinct tasks: binary classification (Original vs. Fake) and multiclass classification into four categories ("FALSE", "HALF TRUE", "MOSTLY FALSE", and "PARTLY FALSE"). The task presents unique challenges due to Malayalam's linguistic complexity and the contextual nuances inherent in social media discourse. The following are the key contributions of this work:

- Development of a Multi-Pooling Feature Fusion (MPFF) architecture incorporating XLM-RoBERTa with CLS-Mean-Max pooling mechanism for robust Malayalam fake news detection
- Extensive experimentation on both binary and multiclass scenarios, establishing the efficacy of the model through rigorous performance metrics and comparative analysis.

The implementation details are publicly available in the GitHub repository:https://github.com/Ayon128/Shared-Task/tree/main/Fake%20News.

2 Related Works

Previous research demonstrates diverse approaches to fake news detection in Dravidian languages. (Rahman et al., 2024) achieved a 0.88 F1 score using a pre-trained Malayalam BERT model. (Tabassum et al., 2024) implemented XLMRoBERTa Base and BERT, reaching an F1 score of 0.87. (M et al., 2024) attained 0.86 using transformer models, while (Farsi et al., 2024) found that fine-tuned MuRIL BERT outperformed other multilingual BERT variants with an equivalent 0.86 F1 score. Transformer-based approaches by (Osama et al., 2024) using XLM-R and mBERT achieved 0.85, and (Tripty et al., 2024) reached 0.84 through

customized preprocessing with m-bert. For multiclass detection, (Kodali and Manukonda, 2024) explored BiLSTM classifiers with custom subword tokenizers, while (Anbalagan et al., 2024a) combined TF-IDF features with LaBSE embeddings in Naive Bayes models. (Anbalagan et al., 2024b) developed the MMFD framework achieving 86% accuracy using Gradient Boosting, and (Majumdar et al., 2021) implemented LSTM with word2vec embeddings, showing high training accuracy (98%) but lower validation accuracy (55%) due to overfitting. (Xing et al., 2024) comparatively analyze Mean, Max, and Weighted Sum pooling mechanisms for BERT and GPT in sentiment analysis, emphasizing task-dependent effectiveness.

3 Task and Dataset Description

The DravidianLangTech 2025 workshop presents a shared task on Fake News Detection in Dravidian Languages (Subramanian et al., 2025). This follows previous editions of similar tasks (Subramanian et al., 2023, 2024).The task is divided into two sub-tasks: Task 1 is binary classification to detect whether a text is "Original" or "Fake", and task 2 is multiclass classification to categorize text into four labels: "FALSE", "HALF TRUE", "MOSTLY FALSE", and "PARTLY FALSE". The datasets contain YouTube comments in the Malayalam language (Devika et al., 2024). As shown in Table 1, for task 1, the dataset consists of 5,091 texts divided into 3,257 texts for training, 815 texts for validation, and 1,019 texts for testing. Similarly, for task 2, the dataset consists of 2,100 texts divided into 1,900 texts for training and 200 texts for testing.

Task	Classes	Train	Dev	Test
Task A	Fake	1,599	406	507
	Original	1,658	409	512
Task B	FALSE	1386	-	100
	HALF TRUE	162	-	37
	MOSTLY FALSE	295	-	56
	PARTLY FALSE	57	-	7

Table 1: Distribution of Malayalam texts for binary (with dev set) and multi-class fake news detection tasks.

4 Methodology

We present an efficient framework for detecting fake news in Dravidian languages focusing on Malayalam text. Figure 2 illustrates an abstract overview of the entire system.

4.1 Preprocessing

We implement a systematic preprocessing pipeline to standardize the Dravidian Fake News dataset. The raw text is subjected to multiple cleaning processes, which includes removing URLs, handling emojis, eliminating hashtags and mentions, and normalizing sequential punctuation. This ensures consistent text representation before feeding into our models.

4.2 Augmentation

To address the severe class imbalance in Task 2, we applied random oversampling using scikit-learn. This was necessary due to the severe imbalance between the majority class (FALSE: 1386 samples) and minority classes, with a focus on "PARTLY FALSE", which has a mere 57 samples. We systematically oversampled minority classes ("HALF TRUE", "MOSTLY FALSE", and "PARTLY FALSE") with replacement to a total count of 200 samples in each category, thus greatly increasing representation across different categories. Without such a balanced sampling strategy, our model would suffer from strong bias towards the "FALSE" class, which forms about 66% of our training set, thus ignoring important subtleties in the severely underrepresented "PARTLY FALSE" category with just below 3% of the total samples. Our results show that such a sampling technique greatly improved minority class classification performance without negatively impacting general accuracy.

4.3 ML-based Approach

We employed several classical machine learning (ML) approaches for Dravidian fake news detection, including logistic regression (LR), support vector machine (SVM), random forest (RF), and XGBoost (XGB). Features were extracted from the preprocessed dataset using the TF-IDF vectorizer (Takenobu, 1994). The RF classifier was configured with 150 estimators and a minimum split threshold of 15 samples. For XGBoost, we utilized a learning rate of 0.05, 150 estimators, and a maximum depth of 6. The SVM classifier was implemented with a radial basis function (RBF) kernel, while Logistic Regression was configured with L2 regularization.

4.4 DL-based Approach

We implemented four deep learning architectures for Dravidian fake news detection, each processing

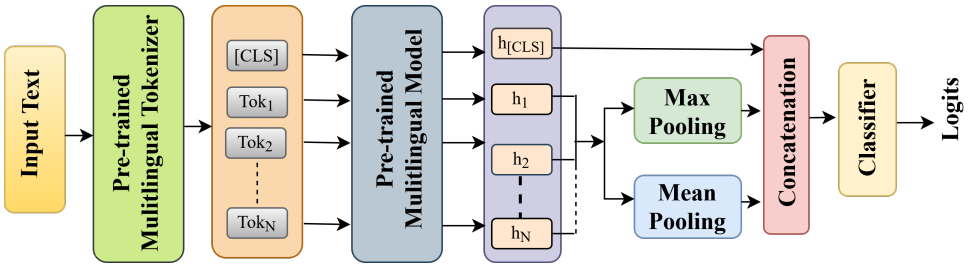


Figure 1: Architecture of the proposed Multi-Pooling Feature Fusion (MPFF) model utilizing XLM-RoBERTa for fake news detection, leveraging [CLS] token and mean and max pooling for enhanced classification.

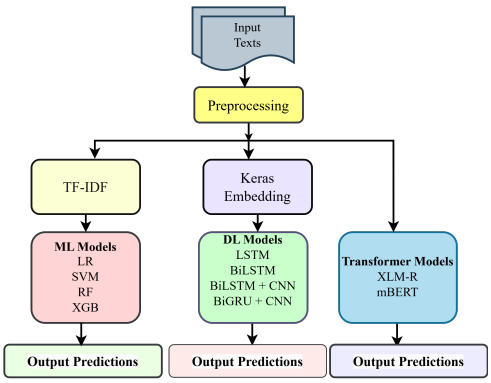


Figure 2: Overview of the Dravidian fake news detection pipeline incorporating traditional ML models with TF-IDF features, Deep Learning architectures using Keras embeddings, and Transformer-based approaches.

text through a 256-dimensional Keras embedding layer. The first architecture employs dual LSTM layers (64 and 32 units) with dropout (0.25) and batch normalization. The second utilizes a bidirectional LSTM with two layers (256 and 128 units). The third combines BiLSTM (256 units) with a dual-stage CNN (128 and 64 filters, kernel size 3), while the fourth substitutes GRU units for LSTM to optimize computational efficiency. We trained all models using Adam optimizer (learning rate 0.001) for 15 epochs with early stopping, implementing binary cross-entropy loss for Task 1 and categorical cross-entropy loss for Task 2.

4.5 Transformer-based Approach

For both tasks, we leveraged two multilingual transformer architectures: mBERT and XLM-RoBERTa (Devlin, 2018; Conneau, 2019), accessed via the Hugging Face platform (Wolf, 2019). The models were implemented in PyTorch with AdamW optimization, using batch size 32 and early stopping across 15 epochs. To ensure model robustness, we implemented k-fold cross-validation strategies: 10 and 7 folds for task 1 and task 2 respectively.

4.6 Multi-Pooling Feature Fusion (MPFF)

Our methodology introduces a Multi-Pooling Feature Fusion (MPFF) architecture, as shown in Figure 1, based on XLM-RoBERTa for both fake news detection tasks. The system processes input text through a pre-trained multilingual tokenizer, i.e., XLM-RoBERTa, generating tokens including a [CLS] token and sequence tokens. These are fed into the pre-trained multilingual model, producing hidden representations for each token. For both binary and multiclass classification tasks, our architecture implements parallel operations: [CLS] token extraction, max pooling, and mean pooling across token representations. These features are concatenated before passing through a classifier layer for final prediction output.

5 Experiments and Results

Table 2 presents the performance analysis across different model architectures. Our methodology introduces a Multi-Pooling Feature Fusion (MPFF) approach, implemented through [CLS] + Mean + Max pooling strategy in transformer architectures. For Task 1 (binary classification), traditional ML models demonstrated consistent performance, achieving macro-averaged F1 scores ranging from 0.735 to 0.775. Among DL approaches, BiLSTM achieved the best performance with F1 score 0.785, while BiGRU + CNN showed comparable results with F1 score 0.783, though BiLSTM + CNN significantly underperformed with F1 score 0.314. The MPFF strategy with XLM-RoBERTa significantly outperformed baseline approaches, achieving optimal results with F1 score 0.874 in Task 1, followed by mBERT with F1 score 0.862. For Task 2's multiclass scenario, our proposed MPFF approach with XLM-RoBERTa demonstrated superior performance with F1 score 0.628, significantly outperforming ML models and DL approaches. These results validate the effec-

Model	Pooling Strategy	Task 1 Performance			Task 2 Performance		
		Pr	Re	F1	Pr	Re	F1
ML Models							
LR	-	0.775	0.775	0.775	0.193	0.254	0.225
SVM	-	0.764	0.764	0.764	0.325	0.254	0.225
RF	-	0.743	0.735	0.735	0.442	0.294	0.294
XGB	-	0.773	0.752	0.752	0.315	0.283	0.283
DL Models							
LSTM	-	0.493	0.493	0.493	0.620	0.620	0.620
BiLSTM	-	0.785	0.785	0.785	0.505	0.505	0.505
BiLSTM + CNN	-	0.314	0.314	0.314	0.610	0.610	0.610
BiGRU + CNN	-	0.783	0.783	0.783	0.035	0.035	0.035
Transformer Models							
mBERT	[CLS]	0.858	0.861	0.860	0.710	0.588	0.608
	[CLS] + Mean + Max	0.866	0.862	0.862	**0.721**	0.592	0.622
XLM-RoBERTa	[CLS]	0.871	0.865	0.868	0.670	0.590	0.610
	[CLS] + Mean + Max	**0.875**	**0.874**	**0.874**	0.685	**0.597**	**0.628**

Table 2: Comparative performance analysis of model architectures on the test sets for Tasks 1 and 2. Pr, Re, and F1 denote macro-averaged precision, recall, and F1-score respectively.

tiveness of our MPFF approach in capturing comprehensive text representations for Malayalam fake news detection across both binary and multiclass scenarios.

6 Error Analysis

Analysis of the confusion matrices shown in Figures 3 and 4 reveals significant error patterns across classification tasks. For Task 1, despite achieving 82.8% accuracy in fake content detection (420 correct predictions), the model exhibits a systematic misclassification tendency where 17.2% of fake instances are incorrectly labeled as original. This error pattern predominantly occurs in texts containing subtle misinformation strategies that blend partial factual elements with deceptive content. In

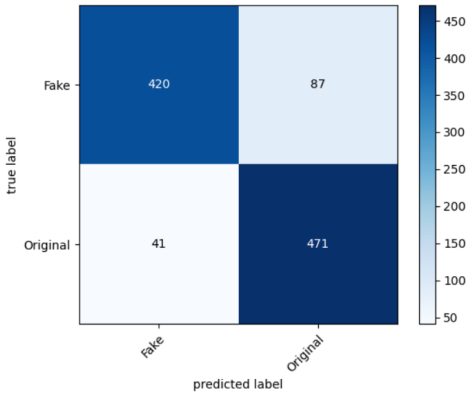


Figure 3: Confusion matrix showing the proposed model's binary classification performance for fake news detection.

the more challenging Task 2, our findings indicate a distinctive hierarchical confusion structure. The model demonstrates robust performance on the "FALSE" category (84% accuracy), yet shows progressive deterioration in performance across the spectrum of partial veracity categories. Specifically, "HALF TRUE" instances manifest confusion with both "FALSE" (12 instances) and "MOSTLY FALSE" (10 instances), while "PARTLY FALSE" classification achieves only 43% accuracy, constituting the model's most significant performance deficiency. Linguistic examination of misclassified samples reveals four primary error sources: contextual dependencies requiring domain-specific knowledge; subtle deceptive linguistic features in partially true content; interference between sentiment markers and factual assessment; and dialectal variations in informal discourse. These findings suggest that while current transformer architectures effectively capture binary veracity distinctions, they remain insufficiently calibrated to the nuanced spectrum of misinformation in low-resource languages like Malayalam.

7 Conclusions

In this study, we conducted a comprehensive analysis of fake news detection in Malayalam YouTube comments through binary and multiclass classification tasks. Our proposed Multi-Pooling Feature Fusion (MPFF) approach with XLM-RoBERTa achieved superior performance through effective integration of [CLS], Mean, and Max pooling features, obtaining macro-average F1 scores of 0.874 and 0.628 for binary and multiclass classification respectively.

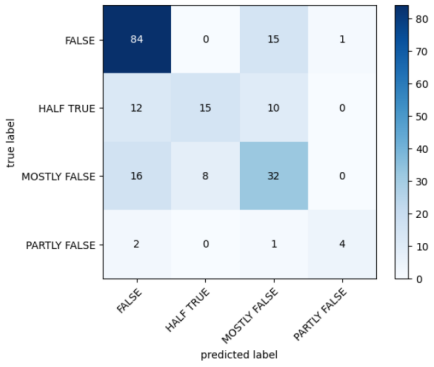


Figure 4: Confusion matrix showing the proposed model's multiclass classification performance for fake news categorization.

8 Limitations

Several limitations emerged in our work. First, despite using 5,091 texts for binary classification, our multiclass dataset of 2,100 texts remained insufficient and imbalanced. This limitation manifested in results, particularly in distinguishing nuanced categories. Second, our model demonstrated weakness in effectively classifying subtle forms of news content. Future work should focus on expanding the Malayalam datasets, exploring advanced architectures for improved classification performance, and investigating advanced Large Language Models (LLMs) which could potentially overcome the dataset limitations and better capture the nuanced distinctions between news categories that our current models struggled with.

References

Akshatha Anbalagan, Priyadharshini T, Niranjana A, Shreedevi Balaji, and Durairaj Thenmozhi. 2024a. WordWizards@DravidianLangTech 2024:fake news detection in Dravidian languages using cross-lingual sentence embeddings. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 162–166, St. Julian's, Malta. Association for Computational Linguistics.

Akshatha Anbalagan, Priyadharshini T, Niranjana A, Shreedevi Balaji, and Durairaj Thenmozhi. 2024b. WordWizards@DravidianLangTech 2024:fake news detection in Dravidian languages using cross-lingual sentence embeddings. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 162–166, St. Julian's, Malta. Association for Computational Linguistics.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Salman Farsi, Asrarul Eusha, Ariful Islam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech EACL2024: Fake news detection in Malayalam language leveraging fine-tuned MuRIL BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 173–179, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Madhumitha M, Kunguma M, Tejashri J, and Jerin Mahibha C. 2024. TechWhiz@DravidianLangTech 2024: Fake news detection using deep learning models. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 200–204, St. Julian's, Malta. Association for Computational Linguistics.

Bhaskar Majumdar, Md. RafiuzzamanBhuiyan, Md. Arid Hasan, Md. Sanzidul Islam, and Sheak Rashed Haider Noori. 2021. Multi class fake news detection using lstm approach. In *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 75–79.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_GoodFellows@DravidianLangTech EACL2024: A transformer-based approach for detecting fake news in Dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192, St. Julian's, Malta. Association for Computational Linguistics.

Tanzim Rahman, Abu Raihan, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek

Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@DravidianLangTech EACL2024: Fake news classification using Malayalam-BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 223–228, St. Julian's, Malta. Association for Computational Linguistics.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_Punctuators@DravidianLangTech-EACL2024: Transformer-based approach for detection and classification of fake news in Malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186, St. Julian's, Malta. Association for Computational Linguistics.

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Zannatul Tripty, Md. Nafis, Antu Chowdhury, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUETSentimentSillies@DravidianLangTech EACL2024: Transformer-based approach for detecting and categorizing fake news in Malayalam language. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 245–251, St. Julian's, Malta. Association for Computational Linguistics.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jinming Xing, Ruilin Xing, and Yan Sun. 2024. Comparative analysis of pooling mechanisms in llms: A sentiment analysis perspective. *arXiv preprint arXiv:2411.14654*.

Mesay Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2024. Habesha@DravidianLangTech 2024: Detecting fake news detection in Dravidian languages using deep learning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 156–161, St. Julian's, Malta. Association for Computational Linguistics.

KCRL@DravidianLangTech 2025: Multi-View Feature Fusion with XLM-R for Tamil Political Sentiment Analysis

Md. Ayon Mia, Fariha Haq, Md. Tanvir Ahammed Shawon, Golam Sarwar Md. Mursalin, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u1804128, u1904051, u1904077}@student.cuet.ac.bd,
sarwarmursalin1015@gmail.com, muhammad_ikhan@cuet.ac.bd

Abstract

Political discourse on social media platforms significantly influences public opinion, necessitating accurate sentiment analysis for understanding societal perspectives. This paper presents a system developed for the shared task of Political Multiclass Sentiment Analysis in Tamil tweets. The task aims to classify tweets into seven distinct sentiment categories: Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral, and None of the above. We propose a Multi-View Feature Fusion (MVFF) architecture that leverages XLM-R with a CLS-Attention-Mean mechanism for sentiment classification. Our experimental results demonstrate the effectiveness of our approach, achieving a macro-average F1-score of 0.37 on the test set and securing the 2nd position in the shared task. Through comprehensive error analysis, we identify specific classification challenges and demonstrate how our model effectively navigates the linguistic complexities of Tamil political discourse while maintaining robust classification performance across multiple sentiment categories.

1 Introduction

Social media platforms have evolved into primary channels for expressing political opinions, generating massive volumes of data that demand sophisticated analysis techniques (Aqlan et al., 2019). While traditional sentiment analysis often emphasizes binary positive-negative classification, contemporary approaches must interpret nuanced evaluative meanings, particularly in political discourse (Alemayehu et al., 2023), (Katta and Hegde, 2019). The growing influence of social media on public opinion formation has highlighted the critical need for advanced sentiment analysis in diverse linguistic contexts, especially for low-resource languages like Tamil. This research addresses this challenge by focusing on Political Multiclass Sentiment Analysis of Tamil tweets, classifying them into seven distinct categories: Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral, and None of the above. The task presents unique challenges due to Tamil's linguistic complexity, the contextual nuances of political discourse, and the inherent informality of social media communication. The key contributions of this work are as follows:

- Development of a Multi-View Feature Fusion (MVFF) architecture incorporating XLM-R with CLS-Attention-Mean mechanism for robust Tamil political sentiment analysis
- Extensive experimental evaluation across multiple sentiment categories, demonstrating the model's effectiveness through rigorous performance metrics and comparative analysis

The following GitHub repository contains the complete implementation details: https://github.com/Ayon128/Shared-Task/tree/main/Political%20Task

2 Related Works

Research in sentiment analysis has demonstrated significant progress across various methodologies. (Nandi and Agrawal, 2016) enhanced sentiment analysis by combining lexical approaches with Linear SVC, achieving 93% accuracy. (Attia et al., 2018) proposed a language-independent CNN model, achieving accuracies of 78.3%, 75.45%, and 67.93% for English, German, and Arabic respectively.(Rao et al., 2020) explored traditional machine learning approaches, where linear kernel SVM reached 80% accuracy. (Derbentsev et al., 2022) compared deep neural networks using Word2vec and Glove vectorization, with CNN achieving 90.1% on IMDb reviews and BiLSTM-CNN reaching 82.1% on Sentiment140. (Alemayehu et al., 2023) compared neural architectures, with CNN-Bi-LSTM achieving 91.60% accuracy. (Dehghani and Yazdanparast, 2023) com-

bined CNN and LSTM architectures, reaching 89% accuracy on their primary dataset. Recent advancements continue to show promising results, with (Ebabu and Chalie, 2024) evaluating models for code-mixed text analysis, where CNN demonstrated superior performance. (Rahman et al., 2024) introduced RoBERTa-BiLSTM, which effectively combines transformer capabilities with bidirectional LSTM networks to capture both contextual embeddings and sequential dependencies, achieving state-of-the-art results on multiple benchmark datasets. These studies demonstrate the effectiveness of hybrid approaches and the importance of appropriate model selection for specific language contexts.

3 Task and Dataset Description

This shared task (Chakravarthi et al., 2025) was organized to analyze political discourse in Tamil language content from X (Twitter) by classifying posts into seven sentiment categories. The objective is to classify Tamil tweets into seven distinct sentiment categories: Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral, and None of the above. Table 1 shows the category-wise distribution of tweets in the training, validation, and test sets. The dataset comprises 5,440 Tamil tweets, distributed across three sets: 4,352 tweets for training, 544 for validation, and 544 for testing. This task aims to advance sentiment analysis capabilities in Tamil, addressing the growing importance of understanding political discourse in low-resource languages on social media platforms.

4 Methodology

Our framework presents an efficient way of carrying out multiclass sentiment analysis of Tamil political comments from X (formerly Twitter). Figure 1 illustrates an abstract overview of the whole system.

4.1 Preprocessing

For standardizing Tamil political tweets, we implement a systematic preprocessing pipeline. The raw text undergoes multiple cleaning operations: URL removal, emoji handling, hashtag and mention elimination, and consecutive punctuation normalization. We also lowercase English text while preserving Tamil script integrity, ensuring consistent text representation for model processing.

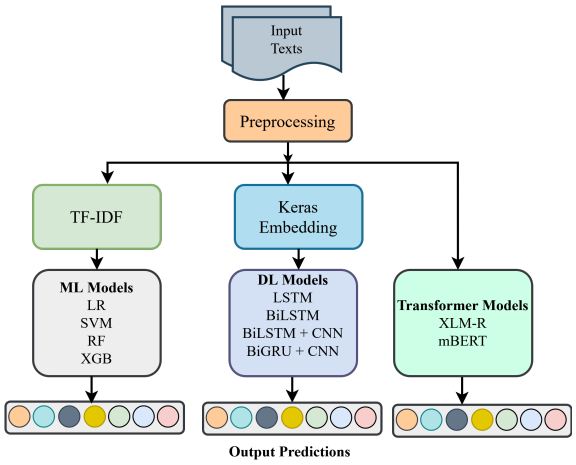


Figure 1: Abstract process of Tamil political sentiment classification using traditional ML models with TF-IDF features, Deep Learning architectures with Keras embeddings, and Transformer models for seven-class sentiment analysis.

4.2 Augmentation

To address the significant class imbalance discovered in Table 1, we employed random oversampling techniques by using scikit-learn. This was crucial due to substantial disparity between the majority classes (Opinionated: 1361 samples) and the minority classes (None: 171 samples). We resampled minority classes (Neutral, Substantiated, Positive, Negative, and None) with replacement specifically to get 200 instances per class. This balanced sampling was essential for preventing model bias towards majority classes and promoting fair learning for all sentiment categories.Without addressing this imbalance, models would likely develop significant bias toward the "Opinionated" class, which constitutes approximately 31% of the training data, while potentially neglecting the "None" category, which represents less than 4% of samples.

4.3 ML-based Approach

For the sentiment classification of Tamil political comments, we implemented traditional ML-based methods: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB). We utilized TF-IDF (Takenobu, 1994) vectorizer to extract features from the preprocessed dataset. The Random Forest classifier was configured with 200 estimators and a minimum split threshold of 10 samples. For XGBoost, we employed a learning rate of 0.1, 100 estimators, maximum depth of 5, and sampling parameters of 0.8 for both instances and features. The SVM

Table 1: Distribution of Tamil tweets across different sentiment categories in the training, validation, and test sets.

Sets	Classes							Total
	Substantiated	Sarcastic	Opinionated	Positive	Negative	Neutral	None	
Train	412	790	1361	575	406	637	171	4352
Val	52	115	153	69	51	84	20	544
Test	51	106	171	75	46	70	25	544

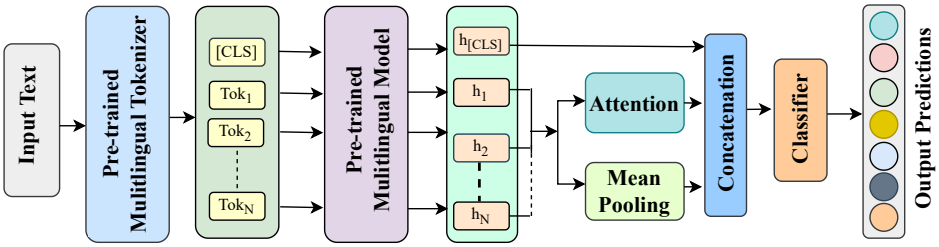


Figure 2: Architecture of XLM-R-based Tamil political sentiment classifier with Multi-View Feature Fusion, combining [CLS], attention, and mean pooling operations for enhanced sentiment classification.

classifier used a linear kernel, while Logistic Regression was implemented with default parameters.

4.4 DL-based Approach

We implemented three different deep learning architectures for the analysis of Tamil political sentiment. A Keras embedding layer was used with a dimensionality of 300 for text representations. The first architecture consists of two LSTM layers (64 and 32 units, respectively) followed by dropout (0.3), batch normalization. The second one proposes a Bidirectional LSTM model that enhances sequential processing. Our third approach implements a hybrid BiLSTM-CNN architecture that combines a Bidirectional LSTM with 128 units, dual-stage CNN with 128 and 64 filters, and kernel size=3. We also developed another variant where LSTM was replaced with GRU units for computational efficiency. All models were trained up to 15 epochs using the categorical cross-entropy loss function with monitoring of validation loss for optimal model selection.

4.5 Transformer-based Approach

We also explored transformer-based approaches by utilizing two multilingual models mBERT (Devlin, 2018) and XLM-R (Conneau, 2019). These pre-trained transformer models were imported from Hugging Face (Wolf, 2019) and implemented using Pytorch library. We fine-tuned both models on the dataset using AdamW optimizer with batch size 32 for 15 epochs, implementing early stopping to prevent overfitting and enhance classification performance.

4.6 Multi-View Feature Fusion (MVFF)

Our methodology introduces a Multi-View Feature Fusion (CLS-Attention-Mean) architecture based on XLM-R for Tamil political sentiment classification. The system processes input Tamil text through a specialized tokenizer, generating a sequence $T = \{[\text{CLS}], \text{tok}_1, \text{tok}_2, \ldots, \text{tok}_n\}$, which is embedded into initial representations $E = \{E_{[\text{CLS}]}, E_1, E_2, \ldots, E_n\}$. The XLM-R transformer processes these embeddings through self-attention mechanisms:

$$H = \text{XLM-R}(E)$$

producing contextualized representations $H = \{h_{[\text{CLS}]}, h_1, h_2, \ldots, h_n\}$. Our architecture implements parallel operations: CLS token extraction $h_{[\text{CLS}]} = H[0]$, mean pooling $h_{\text{mean}} = \text{MeanPool}(h_1 : h_n)$, and the attention mechanism. These features are combined through concatenation:

$$H_{\text{fused}} = [h_{[\text{CLS}]}, h_{\text{att}}, h_{\text{mean}}] .$$

The final classification output is computed through a feed-forward layer with softmax activation:

$$y = \text{softmax}(W \cdot H_{\text{fused}} + b)$$

where y represents the probability distribution over sentiment classes. The complete architecture is illustrated in Figure 2.

Table 2: Performance comparison across different model architectures on the test set, where Pr, Re, and F1 denote macro-averaged precision, recall, and F1-score respectively.

Model	Pooling Strategy	Performance Metric		
		Pr	Re	F1
ML Models				
LR	-	0.36	0.32	0.33
SVM	-	0.36	0.33	0.33
RF	-	0.38	0.33	0.32
XGB	-	0.29	0.31	0.27
DL Models				
LSTM	-	0.35	0.34	0.35
BiLSTM	-	0.32	0.31	0.32
BiLSTM + CNN	-	0.31	0.31	0.31
BiGRU + CNN	-	0.35	0.34	0.34
Transformer Models				
mBERT	[CLS]	0.32	0.31	0.31
	Mean	0.36	0.34	0.34
	Attention	0.34	0.35	0.34
	[CLS] + Mean	0.35	0.32	0.32
	[CLS] + Attention	0.35	0.36	0.35
	[CLS] + Mean + Attention	0.36	0.32	0.33
XLM-R	[CLS]	0.36	0.34	0.33
	Mean	0.37	0.36	0.36
	Attention	0.36	0.35	0.35
	[CLS] + Mean	0.35	0.36	0.35
	[CLS] + Attention	0.35	0.34	0.33
	[CLS] + Mean + Attention	**0.38**	**0.37**	**0.37**

5 Experiments and Results

Table 2 presents the comparative results of different models using macro-averaged precision (Pr), recall (Re), and F1-score (F1). Among ML approaches, RF achieved the highest F1-score of 0.32, slightly outperforming LR and SVM with 0.33, while XGB showed lower effectiveness with 0.27 F1-score. The DL architectures, particularly LSTM and BiGRU+CNN, demonstrated stronger performance with F1-scores of 0.35 and 0.34 respectively. This suggests that the integration of convolutional layers with recurrent architectures enhances feature extraction for Tamil sentiment analysis. Our proposed Multi-View Feature Fusion approach using XLM-R with [CLS]+Mean+Attention strategy achieved the best overall performance with precision of 0.38, recall of 0.37, and F1-score of 0.37, significantly outperforming all baselines. The consistent improvement across different pooling strategies validates the effectiveness of feature fusion for capturing diverse aspects of Tamil political sentiment. The mBERT variants also showed competitive performance, with attention-based pooling providing consistent improvements in F1-scores ranging from 0.31 to 0.35. These results demonstrate that transformer-based architectures with sophisticated pooling strategies are more effective at capturing the nuanced sentiments in Tamil political discourse compared to traditional approaches.

6 Error Analysis

The confusion matrix shown in Figure 3, reveals several key misclassification patterns, providing valuable insights into model behavior and limitations. Our Multi-View Feature Fusion model demonstrates strongest performance in classifying "None of the above" (88%) and "Opinionated" (51%) categories. However, there is a notable tendency to misclassify most categories as "Opinionated," particularly evident in "Neutral" (40%) and "Negative" (50%) content. "Substantiated" content shows dispersed misclassification across categories, primarily confused with "Neutral" (25%) and "Opinionated" (22%). This indicates difficulty in identifying factual content with supporting evidence, possibly due to the complex linguistic markers used in Tamil to denote substantiation. While "Sarcastic" content achieves reasonable classification accuracy (41%), it faces confusion with "Opinionated" (24%), highlighting the challenge of detecting subtle sarcastic cues in text-only political communication. The relatively lower performance on "Neutral" and "Positive" categories (14% and 17% true positive rates respectively) suggests a model bias toward more distinctly marked categories, which could be addressed through balanced training data.

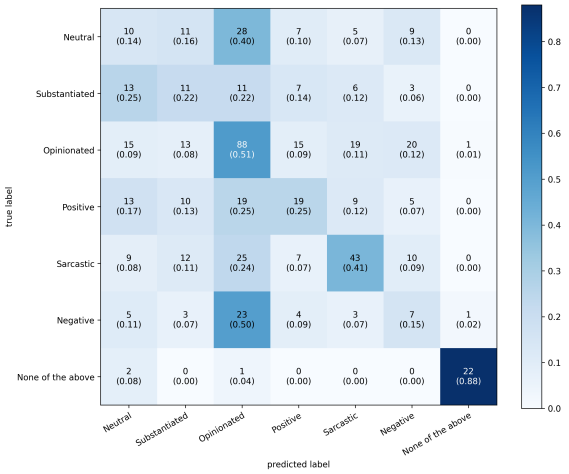


Figure 3: Confusion matrix demonstrating the proposed model's classification performance across seven categories.

7 Conclusions

In this study, we conducted a comprehensive analysis of Tamil political sentiment classification using various machine learning, deep learning, and transformer-based approaches. Our experimental results demonstrated that the Multi-View Feature Fusion approach with XLM-R achieved superior performance through effective integration of [CLS], mean, and attention-based features, obtaining a macro-average F1 score of 0.37. Deep learning architectures like LSTM and BiGRU+CNN showed promising results (F1: 0.35, 0.34), outperforming traditional machine learning approaches, while mBERT variants demonstrated competitive performance with attention-based pooling strategies. The results highlight the importance of combining multiple feature views when analyzing the complex linguistic patterns in Tamil political content, providing a foundation for future sentiment analysis research in low-resource languages.

8 Limitations

Several limitations can be noted in our work. First, the relatively modest F1-scores across all models indicate inherent challenges in Tamil political sentiment analysis. The dataset size constraints and class imbalance issues significantly impacted model development and performance, as evident in our results section. Secondly, our employed models showed limitations in effectively capturing nuanced political sentiments in Tamil text, particularly for complex expressions. The models also underperformed when analyzing tweets without considering broader conversational context or cultural references crucial for accurate classification. Future work should explore advanced techniques for handling class imbalance, larger Tamil political datasets, and enhanced architectures for better sentiment understanding.

References

Fikirte Alemayehu, Million Meshesha, and Jemal Abate. 2023. Amharic political sentiment analysis using deep learning approaches. *Scientific Reports*, 13(1):17982.

Ameen Aqlan, Dr. Manjula Bairam, and R Lakshman Naik. 2019. *A Study of Sentiment Analysis: Concepts, Techniques, and Challenges*, pages 147–162.

Mohammed Attia, Younes Samih, Ali Elkahky, and Laura Kallmeyer. 2018. Multilingual multi-class sentiment classification using convolutional neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Mohammad Dehghani and Zahra Yazdanparast. 2023. Political sentiment analysis of persian tweets using cnn-lstm model. *arXiv preprint arXiv:2307.07740*.

Vasily D Derbentsev, Vitalii S Bezkorovainyi, Andriy V Matviychuk, Oksana M Pomazun, Andrii V Hrabariev, and Alexey M Hostryk. 2022. A comparative study of deep learning models for sentiment analysis of social media texts. In *M3E2-MLPEED*, pages 168–188.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yitayew Ebabu and Minalu Chalie. 2024. Sentiment analysis for amharic-english code-mixed sociopolitical posts using deep learning.

Padmaja Katta and Nagaratna Parameshwar Hegde. 2019. A hybrid adaptive neuro-fuzzy interface and support vector machine based sentiment analysis on political twitter data. *International Journal of Intelligent Engineering & Systems*, 12(1).

Vikash Nandi and Suyash Agrawal. 2016. Political sentiment analysis using hybrid approach. *International Research Journal of Engineering and Technology*, 3(5):1621–1627.

Md Mostafizer Rahman, Ariful Islam Shiplu, Yutaka Watanobe, and Md Ashad Alam. 2024. Robertabilstm: A context-aware hybrid model for sentiment analysis. *arXiv preprint arXiv:2406.00367*.

Dr D Rajeswara Rao, S Usha, S Krishna, M Sai Ramya, G Charan, and U Jeevan. 2020. Result prediction for political parties using twitter sentiment analysis. *International Journal of Computer Engineering and Technology*, 11(4).

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

TensorTalk@DravidianLangTech 2025: Sentiment Analysis in Tamil and Tulu using Logistic Regression and SVM

K. Anishka[1], Anne Jacika J[1]
[1] SSN College of Engineering , Tamil Nadu, India
anishka2310506@ssn.edu.in, annejacika2310581@ssn.edu.in

Abstract

Words are powerful; they shape thoughts that influence actions and reveal emotions. On social media, where billions of people share their opinions daily. Comments are the key to understanding how users feel about a video, an image, or even an idea. But what happens when these comments are messy, riddled with code-mixed language, emojis, and informal text? The challenge becomes even greater when analyzing low-resource languages like Tamil and Tulu. To tackle this, TensorTalk deployed cutting-edge machine learning techniques such as Logistic regression for Tamil language and SVM for Tulu language , to breathe life into unstructured data. By balancing, cleaning, and processing comments, TensorTalk broke through barriers like transliteration and tokenization, unlocking the emotions buried in the language.

1 Introduction

The modern world is rapidly advancing communication and networking technology. Users express opinions in social media comment sections. Therefore, understanding and analyzing the sentiments from these textual data can significantly improve classifying and recommending products and other services by understanding the general public opinion as observed by the authors of (Taboada, 2016). It is also observed that comments are not often in the same language or in its original form i.e, there are slang words, transliterated words, and words in native script as well. Therefore, developing efficient methods to analyze such diverse textual forms of data becomes imperative. Sentiment Analysis is a field of Natural Language Processing (NLP) that focuses on analyzing and interpreting the emotions, opinions, and sentiments expressed in textual data. This generally includes identifying the emotions as positive, negative or neutral based on certain factors. The authors of (Wankhade et al., 2022) interpret sentiment analysis as identifying and extracting subjective information from text using natural language processing and text mining, and discuss methods to complete the given sentiment analysis task and its applications. In this paper, TensorTalk addresses the task of analyzing sentiments of multifaceted Dravidian languages such as Tamil and Tulu, by analyzing textual data obtained across various social media platforms' comment sections. The data that has been analyzed was found to be highly imbalanced, which is as expected of general public opinions and takes varying forms as expressed earlier featuring transliterated words, original text form, slang/dialect words and other language words as well. TensorTalk has thus used the logistic regression and SVM models to address the problem at hand. The proposed solutions for the task: Sentiment Analysis in Tamil and Tulu were presented to, and evaluated by DravidianLangTech@NAACL2025 for the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025) at NAACL 2025 (Durairaj et al., 2025) . In addition to the academic contributions, the proposed solution can help in practical implications as mentioned earlier for product/service recommendation purposes, etc. The detailed analysis of dataset ,data preprocessing, oversampling techniques, vectorization methods, Logistic Regression, SVM model implementations and results are discussed in the subsequent sections. The code for the classification tasks discussed in this paper can be accessed through this link: [1]

2 Related Work

With rapid digital transformation, user-generated content from social networks, blogs, and forums have become a valuable information source. Sentiment analysis, a subfield of Natural Language Pro-

[1] https://github.com/Anishka-K556/Dravidian-Sentiment-Analysis

cessing (NLP), plays a crucial role in understanding public opinions, emotions, and attitudes from such textual data. The authors of (Chakravarthi et al., 2020) remark that comments from social networks do not follow strict rules of grammar, contain more than one language, and are often written in nonnative scripts, thereby making sentiment classification processes much more difficult. In order to effectively analyze the underlying sentiments and prove insights on current trends and various decision-making applications especially in Dravidian languages is in trend according to the authors of (Sambath Kumar et al., 2024). Sentiment analysis in low-resource Dravidian languages like Tamil and Tulu, complicated by code-mixed data, is still in its early stages (Hegde et al., 2023). One of the most significant roles in prediction models is exploration of the data distribution as remarked by the authors of (Bailly et al., 2022) who illustrate the working of models like Logistic Regression and the ways training dataset size and interactions affect the performance of these prediction models. Many significant improvements have been made in the fields of sentiment analysis of textual data in Tamil and Tulu using traditional models like logistic regression as illustrated by authors of (Ponnusamy et al., 2023). According to the authors of (Fang and Zhan, 2015) major components of sentiment analysis include processing the data, vector generation, feature extraction etc. Despite being low resourced languages, the evolution of these Dravidian languages into much more modern forms needs to be understood and researched upon from the perspective of the authors of (Hegde et al., 2022). Datasets obtained from various social media platforms, or in general datasets for various real-time problems are mostly imbalanced. To handle these imbalanced datasets, either oversampling or undersampling techniques are applied. According to the authors of (Elreedy et al., 2024), the SMOTE method generates new synthetic data patterns by performing linear interpolation between minority class samples and their K nearest neighbors, which need not always conform to the original distribution. These can thus, boost a model's performance.Representation of textual data in a form that is compatible with the training models also determines the model's performance. In (Qorib et al., 2023),the authors have extensively tested and implemented various vectorization methods paired with different models to achieve higher performance. Higher performance and model performance generally follow from thoroughly cleaned data. Among multiple vectorization methods, the TF-IDF method known for its robustness and efficiency has been implemented. Textual data are better represented by vectorization methods such as TF-IDF methods that aid the models in training. The authors of (Liu et al., 2018) remark that in addressing problems, such as ignoring contextual semantic links and different vocabulary importance in traditional text classification techniques, vectorization techniques such as word2vec, TF-IDF methods improve traditional and deep learning methods' performances. Vectorization techniques adapted from word2vec such as fasttext embeddings have proven to represent Tamil and Tulu texts better as observed by the authors of (K et al., 2023). Pre trained transformer models have also been used to explore and understand the underlying sentiment in these texts coupled with feature extraction techniques by the authors of (Balaji et al., 2024). Many developments in textual sentiment analysis have been made using deep learning models as illustrated by the authors of (Tang et al., 2015) who have remarked that deep learning approaches emerge as powerful computational models as they discover intricate semantic representations of texts automatically from data without feature engineering.

3 Dataset and Task Description

The problem at hand is to detect and classify the sentiments in textual data obtained from various social platfoms such as comments/posts in Tamil and Tulu languages. Due to the diversity and complexity of the Dravidian languages due to code-mixed data there is a growing demand for improving and developing models for sentiment analysis in these Dravidian languages. The given dataset instances of comments and posts are such that they may contain more than one sentence but the average length is 1. The data that has been supplied is found to have classes such as positive, negative, mixed feelings and unknown state for Tamil data and positive, negative, neutral, mixed and not Tulu for Tulu data and is highly imbalanced, in accordance with real world scenarios. Refer Table 1 that shows the distribution of the textual data that has been provided for the given problem. It is noted that highly imbalanced classes thus lead to biasing of the model i.e. overfitting (latching onto irrelevant patterns) of the majority class and poor generalization of the minority classes.

Task	Labels	No. of instances
Tamil Sentiment Analysis	Positive	18145
	Negative	4151
	Mixed feelings	3662
	Unknown state	5164
Tulu Sentiment Analysis	Positive	3769
	Negative	843
	Mixed feelings	1114
	Neutral	3175
	Not Tulu	4400

Table 1: Dataset Distribution

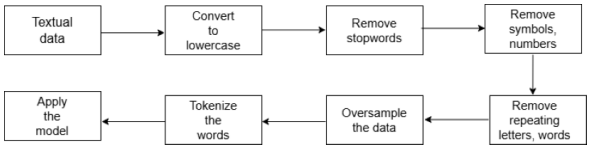


Figure 1: Text Preprocessing and Model Application Workflow

4 Methodology

Given the linguistic diversity and grammatical uniqueness of the language, TensorTalk refined traditional textual analysis methods and applied vectorization techniques. Additionally, TensorTalk has utilized models such as Logistic Regression to effectively identify patterns within the dataset, which was enhanced through vectorization and oversampling. Analyzing the distribution of the given data revealed that the data were highly imbalanced, as observed in the Tamil data set, where positive-labeled data occur predominantly throughout. By creating synthetic samples of the minority class, oversampling increases the representation of the minority class, helping the model learn its patterns. Thus, a better representation of the minority class can be achieved using the SMOTE technique compared to undersampling the majority class, which may lead to data loss and poor generalization, or simple synthetic generation, which can cause overfitting by duplicating existing samples without adding meaningful variation. To tackle this problem, TensorTalk experimented with classical machine learning models, including Logistic Regression and SVM, on textual data that underwent various preprocessing steps along with tokenization methods used to represent text in vectorized forms. These models are implemented along with SMOTE to handle the imbalanced classes in the dataset. These models were chosen due to their robustness. These are discussed in detail below to illustrate their role in improving model training. Refer Figure 1. The decision to choose ML models listed above over deep learning models aligns with the requirements of low computational costs, availability of dataset, need for interpretability, and simplicity. Taking into consideration that the given dataset is on the smaller side TensorTalk decided to opt for traditional models like SVM and Logistic Regression over deep learning models such as LSTM. After adequate testing and hyperparameter tuning, the development data set was used to train the model to improve the model training accuracy.

4.1 Preprocessing

Data generated from most real-time problems contain noise, missing values, physically impossible values, and format that might not be compatible with the models. Thus, to overcome these issues, preprocessing is performed on the data which significantly increases the efficiency of any model. The preprocessing techniques that TensorTalk has used include removing null values and unifying the spaces between words. Considering the fact that comments usually contain various other punctuations, emojis and symbols, these have been removed as well. Some traditional cleaning methods such as removing stopwords ,lemmatization have also been applied. In addition to these, some unique cleaning techniques such as removing duplicate words in an entry, removing repeating letters, timestamps, etc. have also been implemented. To address the issue of class imbalance SMOTE techniques have been applied on the dataset to oversample the minority class. This method has allowed us to generate samples to balance the dataset across all classes. In comparison, to undersampling the majority class, TensorTalk preferred oversampling to avoid the risk of potential loss of information. Finally, the dataset was subject to tokenization using the method of TF-IDF Term Frequency-Inverse Document Frequency. This feature extraction technique allowed us to represent the textual data in a vectorized format by measuring how often a term has appeared in a document. By computing the importance of the term by this weight and reducing the dimensionality, the dataset is converted to a form apt for models to learn from.

4.2 Task: Sentiment Analysis in Tamil

The first task to classify the given textual data was implemented using the Logistic Regression model into multiple classes for varying degrees of sentiment polarity. Logistic Regression, being a linear model, was used to effectively classify the

given textual data into multiple classes in the given problem, where the data had been subject to TF-IDF techniques prior along with the preprocessing techniques mentioned above, by using the logistic/sigmoid function in determining the probability of a given text belonging to a specific class. TensorTalk chose to implement this model for the given problem due to its ability to act as a good baseline model for most text classification tasks. The model proved to be much simpler when compared to the other models, along with its ability to adapt for considerably large datasets, higher interpretability, speed and scalability.

4.3 Task: Sentiment Analysis in Tulu

The task is to classify the given textual data was implemented using the machine learning model Support Vector Machine (SVM). Support Vector Machine (SVM) which is known for it's high dimensional data handling, binary and multi-class classification and also for it's robust to overfitting. This model proved the strength of the SVM in handling complex classification tasks , even in the presence of noisy and unbalanced textual data. TensorTalk chose SVM because one of it's key advantages is it's ability to work well with sparse and high-dimensional feature spaces, which are common in textual data. SVM can capture complex patterns and relationships within text features. It also performs well even with relatively small training datasets, making it a strong choice when labeled data are limited. In this implementation, SVM demonstrated its strength in handling complex classification tasks, even in the presence of noisy and imbalanced data.

4.4 Results and Discussions

The performance of machine learning models such as Logistic Regression and Support Vector Machine (SVM) in sentiment analysis in Tamil language and Tulu language respectively was evaluated on key metrics such as accuracy, precision, recall, and F1 score. The evaluation of the model was mainly based on the macro-averge F1 score, during the training of the model, which was found to be 46% for Tamil and 55% for Tulu, mentioned in table 2. It was found that among all the other models that were trained and tested, the Logistic Regression model gave the best performance, especially when combined with the SMOTE and TF-IDF techniques. The underlying reason could be attributed to the TF-IDF's inverse frequency technique and better generalization of minority classes. In the case of sentiment analysis in Tulu, the SVM model showed better performance compared to other models, which could likely have resulted as SVM performs well in high-dimensional spaces by finding an optimal separating hyperplane. In addition, the test scores for both tasks were found to be 24% for Tamil and 53% for Tulu, refer to table 3. The Logistic regression model's F1 score of 47% indicates that it performed reasonably well on validation data. However, the sharp drop to 24% in the test data suggests that the model struggles to generalize. The few plausible reasons for these could be overfitting of the model, excessive tuning to validation dataset, etc. Although the exact cause remains unknown, TensorTalk believes that better preprocessing techniques could still improve the model's performance. Future work may focus on refining the models by incorporating deep learning models.

Task	Macro Avg F1 score
Tamil Sentiment Analysis	0.46
Tulu Sentiment Analysis	0.55

Table 2: Sentiment Analysis Cross Validation Score

Task	Macro Avg F1 score
Tamil Sentiment Analysis	0.24
Tulu Sentiment Analysis	0.53

Table 3: Sentiment Analysis Test Score

5 Conclusion

As communication continues to grow exponentially, so must techniques and models for machines to analyze them. Through this paper, TensorTalk addresses sentiment analysis in low-resource Dravidian languages, Tamil and Tulu, using classical machine learning. TensorTalk has employed Logistic Regression for Tamil and SVM for Tulu, overcoming challenges such as code-mixed data and transliterations. Our proposal includes preprocessing, tokenization, and oversampling that improve model performance. Future work will explore deep learning and domain-specific embeddings to enhance classification. This research advances Dravidian language processing.

6 Limitations

The proposed models have faced challenges in training to classify the given data into the required labels. This is primarily because low-resource languages, especially in their transliterated form, do not easily conform to specific spellings or distinct word boundaries. As a result, identifying and removing stopwords becomes particularly difficult. The presence of diverse slang and dialectal variations in textual data from social media is found to be less effective than tokenizing text in a standardized language such as English. Due to the complex morphology of Dravidian languages, sentiment analysis becomes more challenging as words can take different forms based on tense, gender, and other grammatical variations. Sentences often mix between English and native languages which makes it difficult to predict labels accurately.The informal nature of social media text introduces noise in the form of misspellings, abbreviations, and emojis, further complicating preprocessing. The same word or phrase can express different sentiments based on context, and traditional models like SVM or Logistic Regression struggle to capture this dynamic contextuality. Support Vector Machine (SVM) and Logistic Regression struggle with contextual and sematic meanings in text, making them less effective. Logistic Regression and SVM are sensitive to the scale of input features, requiring feature normalization for optimal performance. SVM and Logistic Regression can perform poorly when the data is imbalanced, especially in real-world scenarios where some sentiment labels may be underrepresented, as observed in this case. Although oversampling technique has been applied, the model still struggles and the potential reason for this could be overfitting. This emphasizes the need for advanced language models and preprocessing techniques to enhance sentiment analysis in low-resourced languages.

References

Alexandre Bailly, Corentin Blanc, Élie Francis, Thierry Guillotin, Fadi Jamal, Béchara Wakim, and Pascal Roy. 2022. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Computer Methods and Programs in Biomedicine*, 213:106504.

Shreedevi Balaji, Akshatha Anbalagan, Priyadharshini T, Niranjana A, and Durairaj Thenmozhi. 2024. WordWizards@DravidianLangTech 2024: Sentiment analysis in Tamil and Tulu using sentence embedding. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 218–222, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Dina Elreedy, Amir F Atiya, and Firuz Kamalov. 2024. A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning. *Machine Learning*, 113(7):4903–4923.

Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big data*, 2:1–14.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, SUBALALITHA CN, Lavanya S K, Thenmozhi D, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Rachana K, Prajnashree M, Asha Hegde, and H. L Shashirekha. 2023. MUCS@DravidianLangTech2023: Sentiment analysis in code-mixed Tamil and Tulu texts using fastText. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 258–265, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Cai-zhi Liu, Yan-xiu Sheng, Zhi-qiang Wei, and Yong-Quan Yang. 2018. Research of text classification based on improved tf-idf algorithm. In *2018 IEEE international conference of intelligent robotic and control engineering (IRCE)*, pages 218–222. IEEE.

Kishore Kumar Ponnusamy, Charmathi Rajkumar, Prasanna Kumar Kumaresan, Elizabeth Sherly, and Ruba Priyadharshini. 2023. VEL@DravidianLangTech: Sentiment analysis of Tamil and Tulu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 211–216, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Miftahul Qorib, Timothy Oladunni, Max Denis, Esther Ososanya, and Paul Cotae. 2023. Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on covid-19 vaccination twitter dataset. *Expert Systems with Applications*, 212:118715.

Lavanya Sambath Kumar, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 2024. Overview of second shared task on sentiment analysis in code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 62–70, St. Julian's, Malta. Association for Computational Linguistics.

Maite Taboada. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2(1):325–347.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(6):292–303.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

TeamVision@DravidianLangTech 2025: Detecting AI generated product reviews in Dravidian Languages

Shankari S R, Sarumathi P, Bharathi B
Department of Computer Science and Engineering
Sri Sivasubramania Nadar College of Engineering
shankari2210607@ssn.edu.in
sarumathi2210526@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Recent advancements in natural language processing (NLP) have enabled artificial intelligence (AI) models to generate product reviews that are indistinguishable from those written by humans. To address these concerns, this study proposes an effective AI detector model capable of differentiating between AI-generated and human-written product reviews. Our methodology incorporates various machine learning techniques, including Naive Bayes, Random Forest, Logistic Regression, SVM, and deep learning approaches based on the BERT architecture. Our findings reveal that BERT outperforms other models in detecting AI-generated content in both Tamil product reviews and Malayalam product reviews.

1 Introduction

Online product reviews play a vital role in shaping consumer behavior and market dynamics. However, the rise of AI-generated reviews poses a threat to the reliability of online platforms by enabling misleading content. Detecting such reviews is crucial to maintaining consumer trust and informed decision-making. This challenge is amplified in low-resource languages like Malayalam and Tamil, which feature complex linguistic structures and limited annotated datasets. This research focuses on developing models to identify AI-generated reviews in these languages, leveraging both machine learning and deep learning techniques. By addressing this gap, the study contributes to AI content detection, supports linguistic diversity, and enhances trust in digital ecosystems.

[1].

[1] https://github.com/Shankarisr/TeamVision-Detecting-AI-generated-product-reviews.git

2 Related Work

Natural Language Processing (NLP) and machine learning have been widely used to detect AI-generated text. Prova (2023) explored various NLP and machine learning-based approaches to identify synthetic text, emphasizing their effectiveness in distinguishing AI-generated content from human-written text (Prova, 2023).

(Akram, 2023) addressed the growing need for reliable evaluation by developing a multi-domain dataset designed to test state-of-the-art APIs and tools for identifying AI-generated content. Building upon this foundation, our study investigates the effectiveness of AI text detection methods.

Desaire et al. (2023) findings revealed that domain-specific prompts could influence the detectability of AI-generated content, making it more challenging for existing detection models to distinguish between human and synthetic text (Desaire et al., 2023).

Gritsay et al. (2022) examined the effectiveness of AI text detection and emphasized the need for more tokens to improve accuracy (Gritsay et al., 2022). (Shimi et al., 2024) addressed an empirical analysis of language detection in Dravidian languages, focusing on challenges and advancements specific to languages like Tamil, Malayalam, Kannada, and Telugu.

H. B. S. and Rangan (2020) conducted a comprehensive survey on Indian regional language processing, highlighting the challenges and advancements in NLP for languages like Tamil and Malayalam (S. and Rangan, 2020).

Ponnusamy (2023) explored the use of ChatGPT-3 models for Tamil text generation, focusing on how AI models can be leveraged to generate coherent and contextually relevant text in Tamil (Ponnusamy, 2023).

3 Dataset

The goal of this task is to develop a model that can effectively detect AI-generated product reviews in Dravidian languages like Tamil and Malayalam. The dataset used for this purpose is sourced from the Detecting AI-generated product reviews in Dravidian languages, provided by DravidianLangTech@NAACL 2025 (Premjith et al., 2025). The training dataset includes the fields id, data, and label, supporting a supervised learning approach. On the other hand, the testing dataset contains only the id and data, which are used exclusively for making predictions. The dataset descriptions are given in Table 1.

Category	Tamil		Malayalam	
	Train	Test	Train	Test
AI	405	-	400	-
Human	403	-	400	-
Total	808	200	800	100

Table 1: Summary of the training and testing dataset entries for Tamil and Malayalam.

4 Methodology

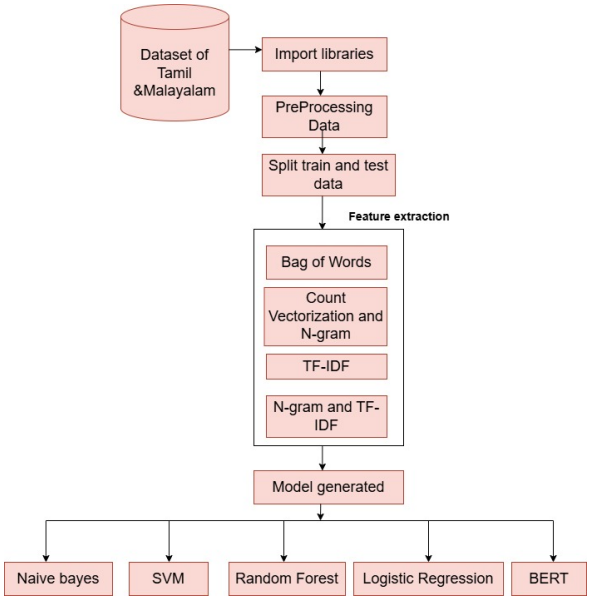


Figure 1: The proposed methodology of the work

4.1 Preprocessing

Text preprocessing is a crucial step in NLP that prepares raw text data for analysis by cleaning and structuring it. It involves several techniques to reduce noise, standardize data, and focus on meaningful content. Stopword Removal filters out common words that carry little semantic meaning, helping to reduce noise and emphasize significant terms. Removing Unwanted characters ensures retaining only relevant content. Tokenization breaks text into smaller units for analysis. Stemming reduces words to their base form by stripping suffixes, simplifying text processing. Lemmatization provides linguistically accurate base forms by grouping words with their root words. This preprocessing pipeline ensures that data is clean, standardized, and optimized for downstream tasks, improving computational efficiency and model accuracy.Figure 2 shows examples of techniques used to reduce noise.

Technique	Tamil Example	Malayalam Example
Stopword Removal	Original: இந்த போன் உண்மையில் மிகவும் நல்லது மற்றும் பயனுள்ளதாக உள்ளது. After: போன் நல்லது பயனுள்ளது.	Original: ഈ ഫോൺ വളരെ നല്ലതാണ്, ഉപകാരപ്രദവുമാണ്. After: ഫോൺ നല്ലത് ഉപകാരപ്രദം.
Removing Unwanted Characters	Original: அற்புதமான தயாரிப்பு!!! பாருங்கள்: www.example.com. After: அற்புதமான தயாரிப்பு பாருங்கள்.	Original: അസാധാരണമായ ഉൽപ്പന്നം!!! പരിശോധിക്കുക: www.example.com. After: അസാധാരണമായ ഉൽപ്പന്നം പരിശോധിക്കുക.
Tokenization	Original: இந்த போன் மிகவும் நல்லது. After: இந்த போன், மிகவும், நல்லது.	Original: ഈ ഫോൺ വളരെ നല്ലതാണ്. After: ഈ, ഫോൺ, വളരെ, നല്ലതാണ്.
Stemming	Original: இருப்பது (being) After:"இரு	Original: ഇരിക്കുന്നു (is) After: ഇരു
Lemmatization	Original: சிறந்த (better) After: சிறந்த	Original: മികച്ച (better) After: മികച്ച

Figure 2: Examples of preprocess techniques

4.2 Feature Engineering

In NLP, feature extraction techniques like BoW, TF-IDF, and n-grams convert text into numerical formats for machine learning. BoW counts word frequency, while TF-IDF weights words based on importance. Using TF-IDF or Count Vectorization with n-grams captures word context. Performance is assessed using F1-Score and accuracy. Figure 1 illustrates the architecture.

4.3 Model Generate

4.3.1 Naïve Bayes

Naïve Bayes, a probabilistic classifier based on Bayes' theorem, assumes feature independence. It achieved 90.12% accuracy (F1: 0.91 AI, 0.88 human) in Tamil and 76.87% accuracy (F1: 0.77 AI, 0.77 human) in Malayalam.

4.3.2 Logistic Regression

Logistic Regression predicts categorical outcomes using probability modeling. It achieved 88.27% accuracy (F1: 0.89 AI, 0.89 human) in Tamil and 76.87% accuracy (F1: 0.75 AI, 0.75 human) in Malayalam.

4.3.3 Support Vector Machine (SVM)

SVM finds the optimal hyperplane for classification. It achieved 89.5% accuracy (F1: 0.89 AI, 0.89 human) in Tamil and 75.62% accuracy (F1: 0.75 AI, 0.76 human) in Malayalam.

4.3.4 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy and reduce overfitting. Achieved 85.8% accuracy (F1: 0.86 AI, 0.91 human) in Tamil and 78.75% accuracy (F1: 0.8 AI, 0.8 human) in Malayalam.

4.3.5 KNN

KNN classifies data based on the majority of nearest neighbors. It achieved 85.8% accuracy (F1: 0.86 for AI human text) in Tamil and 73.12% accuracy (F1: 0.74 for AI, 0.72 for human text) in Malayalam.

4.3.6 BERT

The proposed model uses Multilingual BERT (mBERT) because it supports multiple languages, including Tamil and Malayalam, without requiring separate models for each language.Achieved an accuracy of 90%, with F1 scores of 0.9 for AI and 0.89 for human text.Achieved the highest accuracy of 95%, with F1 scores of 1 for both AI and human text in Malayalam.

4.3.7 Justification

The BERT-based model excelled due to its deep contextual understanding, while Naïve Bayes and Logistic Regression struggled with complex patterns. SVM performed well but was computationally expensive, and Random Forest lacked contextual depth, affecting AI text detection.

5 Results and Discussion

Model	Feature	Prec.	Rec.	F1
NB	BoW	0.88	0.87	0.88
	TF-IDF	0.88	0.87	0.87
	ngram+tf-idf	0.91	0.90	0.89
	CVec + ngram	0.92	0.90	0.90
LR	BoW	0.87	0.87	0.87
	tf-idf	0.87	0.87	0.87
	ngram+tf-idf	0.88	0.88	0.88
	CVec + ngram	0.88	0.88	0.88
SVM	BoW	0.89	0.89	0.89
	ngram+tf-idf	0.90	0.90	0.90
	CVec+ngram	0.88	0.88	0.88
	tf-idf	0.90	0.89	0.89
RF	BoW	0.85	0.85	0.85
	tf-idf	0.86	0.86	0.86
	CVec+ngram	0.84	0.83	0.83
	ngram+tf-idf	0.86	0.86	0.86
DT	BoW	0.83	0.83	0.83
	tf-idf	0.81	0.81	0.81
	ngram + tf-idf	0.82	0.82	0.82
	CVec + ngram	0.85	0.85	0.85
KNN	BoW	0.78	0.63	0.63
	ngram + tf-idf	0.71	0.46	0.44
	CVec + ngram	0.86	0.36	0.36
BERT	**BERT Emb.**	**0.98**	**0.98**	**0.98**

Table 2: Performance Comparison of Models on Tamil Dataset (Premjith et al., 2025)

The performance evaluation of various classifiers on the Tamil and Malayalam shown in Table 2 and Table 3 highlights significant differences in effectiveness across models. BERT emerges as the most accurate classifier, achieving the highest precision, recall, and F1 scores for both AI and human text classification. Specifically, for the Malayalam, BERT reaches an impressive F1 score of 96 for both AI and human text, while in the Tamil, it achieves 97 for AI and 99 for human text. These results emphasize the power of deep learning-based transformer models in understanding complex linguistic patterns, even in low-resource languages. These results suggest that BERT is the go-to choice for classifying AI-human text in languages like Tamil and Malayalam. This highlights a key di-

Model	Feature	Prec.	Rec.	F1
NB	BoW	0.76	0.76	0.76
	tf-idf	0.75	0.75	0.75
	ngram+tf-idf	0.76	0.76	0.76
	CVec+ngram	0.77	0.77	0.77
LR	BoW	0.77	0.76	0.76
	tf-idf	0.76	0.76	0.76
	ngram+tf-idf	0.77	0.77	0.77
	CVec+ngram	0.77	0.77	0.77
SVM	BoW	0.70	0.69	0.69
	tf-idf	0.73	0.73	0.73
	CVec+ngram	0.72	0.72	0.72
	N-gram+tf-idf	0.76	0.76	0.76
RF	BoW	0.78	0.75	0.75
	tf-idf	0.79	0.79	0.79
	CVec+ngram	0.76	0.74	0.74
	N-gram+tf-idf	0.75	0.75	0.75
DT	BoW	0.73	0.72	0.72
	tf-idf	0.69	0.69	0.69
	CVec+ngram	0.74	0.74	0.74
KNN	BoW	0.59	0.52	0.52
	tf-idf	0.74	0.73	0.73
	ngram+tf-idf	0.71	0.71	0.71
	CVec+ngram	0.66	0.39	0.39
BERT	**BERT Emb.**	**0.96**	**0.96**	**0.96**

Table 3: Performance Comparison of Models on Malayalam Dataset (Premjith et al., 2025)

rection for future research: fine-tuning transformer models to better handle low-resource languages, using their strong ability to understand context and generalize across data to boost classification accuracy even further.

5.1 Comparison with Existing AI Text Detection Tools

The proposed BERT-based model surpasses existing AI text detection tools in handling Tamil and Malayalam, while tools like GPTZero, OpenAI AI Text Classifier, GLTR, and Turnitin primarily focus on English. Unlike statistical or probabilistic models, the proposed approach leverages TF-IDF, n-grams, and contextual embeddings, allowing better customization and adaptability for low-resource languages. While existing tools lack fine-tuning for non-English texts, the proposed model effectively detects AI-generated reviews with moderate explainability, making it more suitable for review detection in underrepresented languages compared to general-purpose detection tools.

6 Conclusions

The experimental results on Malayalam and Tamil datasets demonstrate that transformer-based models, particularly BERT, significantly outperform traditional machine learning approaches in classification accuracy. BERT achieves the highest precision, recall, and F1 scores across both datasets, reinforcing its effectiveness in handling complex linguistic structures in low-resource languages. While traditional classifiers like Naïve Bayes, Logistic Regression, SVM, and Random Forest show moderate performance, models like Decision Tree and KNN struggle to generalize effectively. These findings highlight the importance of leveraging deep learning models for AI-human text classification, ensuring reliable detection methods in the face of rapidly advancing AI-generated content.

7 Limitations

Tamil and Malayalam lack large, high-quality datasets, making AI models prone to bias and inaccuracies.Many users blend Tamil/Malayalam with English or use Romanized script, which traditional models struggle to process. Rich morphology in these languages makes tokenization and feature extraction difficult, reducing model accuracy. AI models, including BERT, often misinterpret sarcasm and subtle sentiments, leading to errors.Training BERT for Tamil and Malayalam requires significant resources, limiting practical use.Language and user reviews evolve over time, causing model degradation. Continuous updates and retraining are essential to maintain classifier accuracy in real-world applications.

8 Error Analysis

False Positives: Formal or repetitive genuine reviews misclassified. False Negatives: AI-generated reviews mimicking humans went undetected. Language Issues: Struggled with code-mixed text and dialects. Improvements: Train on diverse data, refine code-mixed handling, and add context-aware features.

References

Arslan Akram. 2023. An empirical study of ai generated text detection tools. *ArXiv*, abs/2310.01423.

Heather Desaire, Andrea E. Chua, Min Gyu Kim, and David Hua. 2023. Accurately detecting ai text when chatgpt is told to write like a chemist. *Cell Reports Physical Science*, 4(11):101672.

German Gritsay, Andrey Grabovoy, and Yu. V. Chekhovich. 2022. Automatic detection of machine generated texts: Need more tokens. *2022 Ivannikov Memorial Workshop (IVMEM)*, pages 20–26.

R. Ponnusamy. 2023. Tamil text generation using chatgpt-3 models. [Online]. Available: ponnusamyr@citchennai.net.

B Premjith, Nandhini K, Bharathi Raja Chakravarthi, Thenmozhi Durairaj, Balasubramanian Palani, and Kumaresan Prasanna Kumar Thavareesan, Sajeetha. 2025. Overview of the Shared Task on Detecting AI Generated Product Reviews in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

N. N. I. Prova. 2023. Detecting ai generated text based on nlp and machine learning approaches. [Online]. Available: nuzhatnsu@gmail.com.

H. B. S. and R. K. Rangan. 2020. A comprehensive survey on indian regional language processing. *SN Applied Sciences*, 2(7):1–16.

G Shimi, CJ Mahibha, and D Thenmozhi. 2024. An empirical analysis of language detection in dravidian languages. *Indian Journal of Science and Technology*, 17(15):1515–1526.

CIC-NLP@DravidianLangTech 2025: Fake News Detection in Dravidian Languages

Tewodros Achamaleh[1], Nida Hafeez[1], Mikiyas Mebiratu[2]
Fatima Uroosa[1], Grigori Sidorov[1]
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico
[2]Wolkite University, Department of Information Technology, Wolkite, Ethiopia

Abstract

Misinformation is a growing problem for technology companies and society. Although there is a large body of related work on identifying fake news in predominantly resource languages, there is a lack of such studies in low-resource languages (LRLs). Because corpora and annotated data are scarce in LRLs, the identification of false information remains in the exploratory stage. Fake news detection is critical in this digital era to avoid the spread of misleading information. This study presents an approach to Detect Fake News in Dravidian Languages. Our team CIC-NLP work primarily targets Task 1, which involves identifying whether a given social platform news is original or fake. For the fake news detection problem, we used the mBERT model and utilized the dataset provided by the organizers of the workshop. In this section, we describe our findings and the results of the proposed method. The mBERT model achieved an F1 score of 0.853. The source code is available on GitHub.[1]

1 Introduction

Social media plays an important role in how people convey and access information in the modern world (Omar and Ondimu, 2024). While instant communication and global connectivity offer numerous benefits, they also bring a significant downside: the problem of disseminating fake news at a very fast rate (Schmidt and Wiegand, 2017). Fake news, described as actual falsehoods disinformation that is passed as factual news, is normally spread to influence people, create controversy, or achieve certain objectives (Bharathi et al., 2021). Fake news is the most rampant problem on social media, significantly transforming or even eroding society in terms of truth and democratic values. Because conspiracy theories can be spread on social media platforms within minutes, millions of people constantly receive distorted information telling the truth from the lie. This process destroys the credibility of institutions and the media, which poses a significant problem for the stable development of society and making conscious decisions (Balaji et al., 2023).

The fake news detection problem has become a critical issue, and it is difficult to ensure the integrity of news content (Aïmeur et al., 2023; Subramanian et al., 2023). Although researchers have conducted a lot of work in numerous domains and languages (Abiola et al., 2025b,a; Mehak et al., 2025), FND in Dravidian languages still faces challenges due to its cultural and linguistic characteristics (Anbalagan et al., 2024). Fake news has caused many incidents, such as the Brexit vote in Britain, the 2017 US elections, and the case in South Africa targeting Finance Minister Pravin Gordhan and media professionals. Such updates on social networks can be attributed to rumors or fiction due to existing fake news or hearsay accounts (Shu et al., 2017). Therefore, it is imperative to develop proper methods for detecting fake news (Oshikawa et al., 2018). The challenge of identifying fake news in Dravidian languages demands specific approaches because these languages possess distinct linguistic and cultural traits (Chakravarthi et al., 2021; Arif et al., 2022; Malliga et al., 2023).

In response to these challenges, workshop organizers launched fake news detection task (Subramanian et al., 2025). The shared task presents developers with a new opportunity to advance research and build better systems for detecting fake news across Dravidian languages. By working together researchers and practitioners build new methods to identify fake news in different languages as shared tasks reveal the motivation behind teamwork and innovation. Our team CIC-NLP used mBERT to classify news into real and fake categories. To identify fake news in Dravidian languages, this work investigates the methods, datasets used and demon-

[1]https://github.com/teddymas95/Fake-News-Detection

strate results. Extending from the extant literature, it is our intention to contribute by culturally informed approaches to improve the efficiency of this task.

2 Literature Review

During past years, by exploring and investigating social media platforms, researchers achieved major progress in fake news detection (Bala and Krishnamurthy, 2023; Chen et al., 2023; Subramanian et al., 2025). Normal method to evaluate a post authenticity is to examine different key indicators such as shares, likes and followers etc. Researchers (Rodríguez and Iglesias, 2019) utilized traditional ML approaches like support vector machines and classification trees to identify or examine key indicators. In the same way, other researchers (Shu et al., 2017; Raja et al., 2023) examined user data including social networks and text information through ML tools to extract network and profile attributes.

Researchers (Singh et al., 2018; Sharma et al., 2019, 2020) examined and reviewed how different Deep Learning (DL), Artificial Intelligence (AI) and Machine Learning (ML) methods help to identify fake news. In another work researchers (Granskogen, 2018; Chauhan and Palivela, 2021; Tonja et al., 2022; Tash et al., 2022; Yigezu et al., 2023a,b; Bade et al., 2024; Mersha et al., 2024) examined how NLP and ML approaches help to detect fake news by studying user sentiments and how content is written. Transfer learning and DL model applications have also been considered (Raja et al., 2023). CNNs and RNNs approaches for FND were used in the study by (Goldani et al., 2021). Static and dynamic searches have been analyzed by (Ahmad et al., 2020) where a two-tier ensemble approach to determine the authenticity of websites was conducted using Naive Bayes (Granik and Mesyura, 2017), Random Forest, and Logistic Regression algorithms. Reacher's used the Naïve Bayes classifier which is a Bayesian solution for detecting fake news (Adiba et al., 2020; Subramanian et al., 2024; Devika et al., 2024).

As the attention has shifted towards the use of graph networks, few works have aimed toward developing graph-based approaches. Researchers (Nguyen et al., 2020) presented FANG for novel graphical social context representation and learning. For realistic scenarios, authors (Lu and Li, 2020) designed Graph aware Co-attention Network (GCAN) to identify the authenticity of the source tweet and provide justification through identification of suspicious re-tweeters and important textual residues. (Mahabub, 2020) This work has also been done towards fake news detection in multilingual and low-resource settings using enable classifiers.

Researchers (Lucas et al., 2022) concentrated on COVID-19 misinformation in the Caribbean regions and trained the models in high-resource languages, then transferred the knowledge to low-resource datasets in English, Spanish, and Haitian French. Researchers (Sivanaiah et al., 2022) also developed fake news datasets for low-resource languages such as Malayalam, Kannada, Tamil and Gujarati. For the first time, researchers presented a multilingual and multi-domain fake news detection dataset that spanned over 5 languages and 7 domains and developed a novel BERT-based multi-language and multi-domain fake news detection framework (De et al., 2021). Some previous work has focused on the application of transfer learning and contextual word embedding's in FND (Akram and Shahzad, 2021; Yigezu et al., 2024). (Kalraa et al., 2021) used the transformer-based models for FND in Urdu, and (Ameer et al., 2021) used TL with the BERT model. Authors (Lina et al., 2020) further used CharCNN and RoBERTa to gain word and character-level sentence embedding's where label smoothing was incorporated to enhance generalization capability. In this context, researchers (Palani and Elango, 2023) employed contextual word embedding's with BERT and Roberta language corpora for detecting the fake news in Dravidian languages. In the same way, (Chakravarthi et al., 2022) used feed-forward networks (FFN) with RoBERTa to extract contextually dependent features for fake news detection.

Recent research in multilingual NLP Muhammad et al. (2023) has explored various techniques for text classification in low-resource languages. Muhammad et al. (2025) introduced a transformer-based approach that improved sentiment analysis, which we extend to enhance fake news detection in Dravidian languages by leveraging advanced modeling techniques. To sum up, the existing methods give useful information about how effectively different strategies function in identifying fake news. Still, there is a lack of further research to address new challenges that appear from time to time, multilingual recognition, and other difficulties in effectively detecting fake news in a varied and resource-constrained environment.

Model	Precision	Recall	F1-Score	Accuracy
LSTM	0.2509	0.5000	0.3342	0.5018
DistilBERT (base-uncased)	0.7420	0.7359	0.7345	0.7362
XLM-RoBERTa (base)	0.8350	0.8348	0.8348	0.8348
BERT (base-multilingual-cased)	0.8622	0.8613	**0.8612**	0.8614

Table 1: Model Comparison on the Development Dataset

3 Fake News Detection

3.1 Dataset Analysis

On the shared task fake news detection organized by DravidianLangTech@NAACL 2025, there are datasets for identifying fake vs original social media posts in Malayalam. The datasets we worked with consisted of "Label" and "Text". Each column was divided into development, training, and test subsets, and analysis was conducted as part of our team "CIC-NLP". We have 3,257 samples in our training dataset, 1,599 of which are tagged as 'fake' and 1,658 tagged as 'original,' forming a near balance. Also, the development dataset has 815 samples, including 406 'fake' and 408 'original,' where the evaluation is balanced. The testing dataset is provided with no labels for blind evaluation of model performance. Both datasets have a well-balanced structure so that training and evaluation will be unbiased. The training set contains lots of data for robust model training, while the development set can be used for tuning and validation. The "Text" column, though, may contain noisy or ambiguous content from social media, which introduces preprocessing challenges. The given data provides a solid foundation for building an effective fake news classifier, and we aim to build a robust classifier using this data.

3.2 BERT-base Multilingual Cased Model

The shared task has selected the Bert-base multilingual cased model, a powerful transformer-based language model that supports more than 100 languages, including Malayalam. This model effectively handles linguistic diversity, including transliteration and code-mixing, making it well-suited for multilingual-based applications. It can handle various languages used in the posts. By preventing text casing, the model's tokenizer helps us distinguish proper nouns and context-dependent entities required for identifying fake news. The model is fine-tuned on a task-specific dataset and classifies posts as either "fake" or "original" by using a contextual understanding of how language patterns manifest. We show that it effectively handles noisy social media data like emojis, abbreviations, and grammatical errors and extracts meaningful insights to guarantee accurate classification. By leveraging the BERT-base multilingual model, the CIC-NLP team aims to attain robust performance in fake news detection in Malayalam posts.

4 System Setup and Experiments

4.1 System Setup

For the Malayalam FND Task 1, we use the BERT-base multilingual cased model, fine-tuned on the provided datasets. The setup requires a GPU-enabled environment with libraries such as PyTorch, Hugging Face Transformers, and Scikit Learn for data processing, training, and evaluation. Preprocessing datasets involved mapping labels (fake and original) and tokenizing with the model's tokenizer with padding and truncation. We fine-tune the model with a learning rate of 3e-5, batch size of 16, and five iterations. Evaluation is done on metrics such as accuracy and F1 scores. We evaluate the performance of the development dataset and save predictions for the test dataset. Classification performance is explored using visualization tools such as confusion matrices and ROC curves. It can be fine-tuned efficiently and is good at detecting fake news in Malayalam posts.

4.2 Experiments

We studied the system's performance through experiments with a BERT-base multilingual cased model for FND in Malayalam. The training dataset

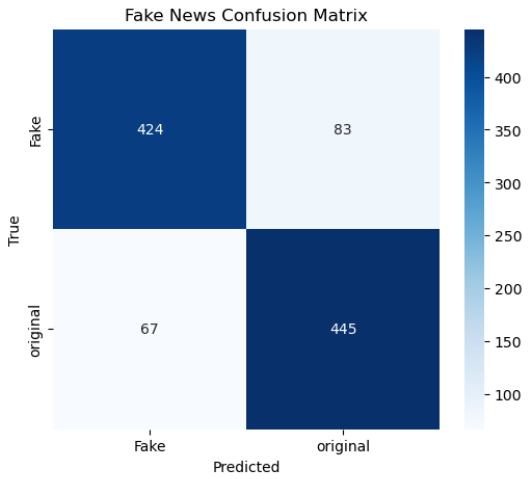


Figure 1: Confusion matrix

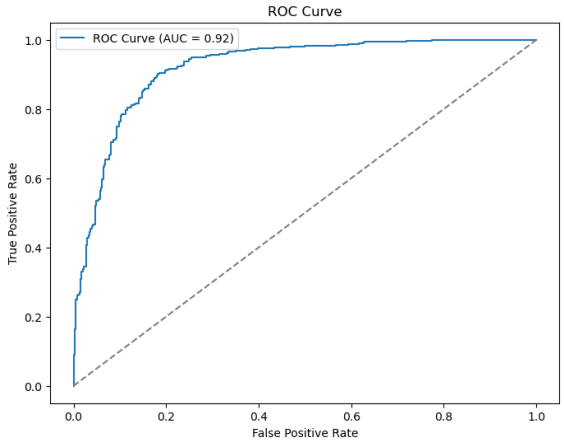


Figure 2: ROC curve

is 3,257 elements, split into 1,599 "fake" and 1,658 "original" labels, and the development dataset contains 815 elements (406 "fake" and 408 "original" labels). For this, we fine-tuned the model using the Hugging Face API Trainer with a learning rate of 3e-5, batch size of 16 and 5 epochs, and weight decay of 0.01. The best checkpoint was saved based on the macro F1 score and evaluations done at the end of each epoch. Accuracy, macro F1 score, and weighted F1 score were used to evaluate the model performance comprehensively. On the development dataset, the fine-tuned model achieved high accuracy and F1 scores for classification between the two classes, "fake" and "original" news posts. A confusion matrix was plotted to visualize the true versus predicted labels, and a ROC curve was plotted to see the contrast and the AUC score from the model to the separation between its classes. In these experiments, we demonstrate that the BERT-base multilingual cased model performs well on fake news detection in Malayalam when using multilingual and noisy data. The results are presented in Table 1 along with Figures 1 and 2, which illustrate the confusion matrix and ROC curve plots.

5 Results

We evaluate the fine-tuned BERT-base multilingual case model on a Malayalam dataset for binary classification. The model showed strong predictive ability on the development set, achieving a macro average F1 score of 0.85, which suggests balanced performance over the two classes, "fake" and "original." The results also demonstrate that the model can generalize well with noisy and relatively imbalanced data. We found that the multilingual BERT architecture is robust to linearity, allowing us to achieve strong performance on Malayalam social media text with negligible performance degradation. The findings highlight the model's effectiveness in detecting fake news in a multilingual environment. The result obtained from the test set is 0.85.

6 Discussion

BERT-multilingual performed best in FND, achieving the highest F1-score, while XLM-RoBERTa also showed strong results. LSTM struggled due to its limited contextual understanding. Class imbalance in Malayalam affected performance, highlighting the need for balanced datasets. The results reinforce the effectiveness of transformer models for low-resource languages. BERT-multilingual outperformed LSTM and DistilBERT, achieving the highest precision and recall. DistilBERT misclassified many fake news samples, and LSTM had difficulty capturing complex linguistic patterns. Future improvements could include domain-specific fine-tuning and metadata integration for better accuracy. Table 1 compares the models.

6.1 Error Analysis

Most misclassifications occurred in the imbalanced Malayalam dataset. False positives were common when fake news closely resembled original content, while false negatives resulted from ambiguous writing styles. Addressing these issues requires better feature representation and fine-tuning. This analysis compares model predictions with the ground truth to identify misclassification patterns and assess overall model performance. Figure 1 shows the confusion matrix.

7 Conclusion

Our team assessed the ability of transformer-based models to identify fake news information in the Malayalam language. Despite the unequal distribution of classes within the dataset, the mBERT-multilingual model achieved the highest accuracy along with the F1-score among other models. The model showed resilient behavior when detecting fake news despite its reduced performance because of the unbalanced dataset. Classification errors stemmed from counterfeit and genuine news pairs that appeared similarly or used ambiguous writing styles. All models experienced difficulties processing uncertain cases because their linguistic feature integration required further enhancement. Future development of fake news detection systems should address enhancements in feature representation combined with domain-based tweaking of models alongside metadata advancement for better identification in low-resource languages.

8 Limitations

This study faced several challenges, primarily due to class imbalance affecting the performance of the Malayalam model. The model struggled with generalization as the dataset lacked sufficient diversity in writing styles. Misclassification issues arose when fake news closely resembled original content, making detection difficult. Additionally, the models had trouble processing ambiguous cases with misleading linguistic patterns. Future improvements should focus on expanding the dataset with balanced class distribution and integrating advanced linguistic features to enhance detection accuracy.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Oluwatobi Joseph Abiola, Temitope Olasunkanmi Oladepo, Olumide Ebenezer Ojo, Grigori Sidorov, and Olga Kolesnikova. 2025a. CIC-NLP at GenAI detection task 1: Leveraging DistilBERT for detecting machine-generated text in English. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 271–277, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide Ebenezer Ojo. 2025b. CIC-NLP at GenAI detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 262–270, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Farzana Islam Adiba, Tahmina Islam, M Shamim Kaiser, Mufti Mahmud, and Muhammad Arifur Rahman. 2020. Effect of corpora on classification of fake news using naive bayes classifier. *International Journal of Automation, Artificial Intelligence and Machine Learning*, 1(1):80–92.

Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. 2020. Fake news detection using machine learning ensemble methods. *Complexity*, 2020(1):8885861.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Hammad Akram and Khurram Shahzad. 2021. Ensembling machine learning models for urdu fake news detection. In *FIRE (Working Notes)*, pages 1142–1149.

Iqra Ameer, Claudia Porto Capetillo, Helena Gómez-Adorno, and Grigori Sidorov. 2021. Automatic fake news detection in urdu language using transformers. In *FIRE (Working Notes)*, pages 1127–1134.

Akshatha Anbalagan, T Priyadharshini, A Niranjana, Shreedevi Balaji, and Durairaj Thenmozhi. 2024. Wordwizards@ dravidianlangtech 2024: Fake news detection in dravidian languages using cross-lingual sentence embeddings. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 162–166.

Muhammad Arif, Atnafu Lambebo Tonja, Iqra Ameer, Olga Kolesnikova, Alexander F Gelbukh, Grigori Sidorov, and Abdul Gafar Manuel Meque. 2022. Cic at checkthat!-2022: Multi-class and cross-lingual fake news detection. In *CLEF (Working Notes)*, pages 434–443.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024. Social media fake news classification using machine learning algorithm. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 24–29.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. Abhipaw@ dravidianlangtech: Fake news detection in dravidian languages using multilingual bert. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–238.

Varsha Balaji, B Bharathi, et al. 2023. Nlp_ssn_cse@ dravidianlangtech: Fake news detection in dravidian languages using transformer models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–139.

B Bharathi et al. 2021. Ssncse_nlp@ dravidianlangtech-eacl2021: Offensive language identification on multilingual code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 313–318.

Bharathi Raja Chakravarthi, Mihaela Găman, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, et al. 2021. Findings of the vardial evaluation campaign 2021. In *EACL| VarDial*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John Philip McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, et al. 2022. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the second workshop on language technology for equality, diversity and inclusion*, pages 378–388.

Tavishee Chauhan and Hemant Palivela. 2021. Optimization and improvement of fake news detection using deep learning approaches for societal benefit. *International Journal of Information Management Data Insights*, 1(2):100051.

Sijing Chen, Lu Xiao, and Akit Kumar. 2023. Spread of misinformation on social media: What contributes to it and how to combat it. *Computers in Human Behavior*, 141:107643.

Arkadipta De, Dibyanayan Bandyopadhyay, Baban Gain, and Asif Ekbal. 2021. A transformer-based approach to multilingual fake news detection in low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–20.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Mohammad Hadi Goldani, Reza Safabakhsh, and Saeedeh Momtazi. 2021. Convolutional neural network with margin loss for fake news detection. *Information Processing & Management*, 58(1):102418.

Mykhailo Granik and Volodymyr Mesyura. 2017. Fake news detection using naive bayes classifier. In *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*, pages 900–903. IEEE.

Torstein Granskogen. 2018. Automatic detection of fake news in social media using contextual information. Master's thesis, NTNU.

Sakshi Kalraa, Preetika Vermaa, Yashvardhan Sharma, and Gajendra Singh Chauhan. 2021. Ensembling of various transformer based models for the fake news detection task in the urdu language. In *FIRE (Working Notes)*, pages 1175–1181.

Nankai Lina, Sihui Fua, and Shengyi Jianga. 2020. Fake news detection in the urdu language using charcnn-roberta. *Health*, 100:100.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*.

Jason Lucas, Limeng Cui, Thai Le, and Dongwon Lee. 2022. Detecting false claims in low-resource regions: A case study of caribbean islands. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 95–102.

Atik Mahabub. 2020. A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers. *SN Applied Sciences*, 2(4):525.

S Malliga, Bharathi Raja Chakravarthi, SV Kogilavani, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, and Muskaan Singh. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 59–63.

Gull Mehak, Amna Qasim, Abdul Gafar Manuel Meque, Nisar Hussain, Grigori Sidorov, and Alexander Gelbukh. 2025. TechExperts(IPN) at GenAI detection task 1: Detecting AI-generated text in English and multilingual contexts. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 161–165, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Melkamu Abay Mersha, Girma Yohannis Bade, Jugal Kalita, Olga Kolesnikova, Alexander Gelbukh, et al. 2024. Ethio-fake: Cutting-edge approaches to combat fake news in under-resourced languages using

explainable ai. *Procedia Computer Science*, 244:133–142.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Nelson Odhiambo Onyango, Lilian DA Wanzare, Samuel Rutunda, Lukman Jibril Aliyu, et al. 2025. Afrihate: A multilingual collection of hate speech and abusive language datasets for african languages. *arXiv preprint arXiv:2501.08284*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M Mohammad, Sebastian Ruder, et al. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages. *arXiv preprint arXiv:2302.08956*.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1165–1174.

AS Omar and KO Ondimu. 2024. The impact of social media on society: A systematic literature review. *The International Journal of Engineering and Science*, 13(6):96–106.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Balasubramanian Palani and Sivasankar Elango. 2023. Bbc-fnd: An ensemble of deep learning framework for textual fake news detection. *Computers and Electrical Engineering*, 110:108866.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Álvaro Ibrain Rodríguez and Lara Lloret Iglesias. 2019. Fake news detection using deep learning. *arXiv preprint arXiv:1910.03496*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42.

Uma Sharma, Sidarth Saran, and Shankar M Patil. 2020. Fake news detection using machine learning algorithms. *International Journal of creative research thoughts (IJCRT)*, 8(6):509–518.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Naman Singh, Tushar Sharma, Abha Thakral, and Tanupriya Choudhury. 2018. Detection of fake profile in online social networks using machine learning. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 231–234. IEEE.

Rajalakshmi Sivanaiah, Nishaanth Ramanathan, Shajith Hameed, Rahul Rajagopalan, Angel Deborah Suseelan, and Mirnalinee Thanka Nadar Thanagathai. 2022. Fake news detection in low-resource languages. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 324–331. Springer.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

M Shahiki Tash, Z Ahani, Al Tonja, M Gemeda, N Hussain, and O Kolesnikova. 2022. Word level language identification in code-mixed kannada-english texts using traditional machine learning algorithms. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 25–28.

Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbuk. 2022. Transformer-based model for word level language identification in code-mixed kannada-english texts. *arXiv preprint arXiv:2211.14459*.

Mesay Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2024. Habesha@DravidianLangTech 2024: Detecting fake news detection in Dravidian languages using deep learning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 156–161, St. Julian's, Malta. Association for Computational Linguistics.

Mesay Gemeda Yigezu, Tadesse Kebede, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023a. Habesha@ dravidianlangtech: Utilizing deep and transfer learning approaches for sentiment analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 239–243.

Mesay Gemeda Yigezu, Moges Ahmed Mehamed, Olga Kolesnikova, Tadesse Kebede Guge, Alexander Gelbukh, and Grigori Sidorov. 2023b. Evaluating the effectiveness of hybrid features in fake news detection on social media. In *2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 171–175. IEEE.

CoreFour_IIITK@DravidianLangTech 2025: Abusive Content Detection Against Women Using Machine Learning And Deep Learning Models

Varun Balaji S[1], Bojja Revanth Reddy[1], Vyshnavi Reddy Battula[1], Suraj Nagunuri[1], Balasubramanian Palani[2]

[1]Department of Computer Science and Engineering, IIIT Kottayam, Kerala, India

[2]Assistant Professor, Indian Institute of Information Technology Kottayam

{varun22bcs152, revanth22bcs210, vyshnavi22bcs133, suraj22bcy35, pbala}@iiitkottayam.ac.in

Abstract

The rise in utilizing social media platforms increased user-generated content significantly, including negative comments about women in Tamil and Malayalam. While these platforms encourage communication and engagement, they also become a medium for the spread of abusive language, which poses challenges to maintaining a safe online environment for women. Prevention of usage of abusive content against women as much as possible is the main issue focused in the research. This research focuses on detecting abusive language against women in Tamil and Malayalam social media comments using computational models, such as Logistic regression model, Support vector machines (SVM) model, Random forest model, multilingual BERT model, XLM-Roberta model, and IndicBERT (Rajiakodi et al., 2025). These models were trained and tested on a specifically curated dataset containing labeled comments in both languages. Among all the approaches, IndicBERT achieved a highest macro F1-score of 0.75. The findings emphasize the significance of employing a combination of traditional and advanced computational techniques to address challenges in Abusive Content Detection (ACD) specific to regional languages.

1 Introduction

Detecting abusive content targeting women on Online Social Networks (OSNs) presents a significant challenge in the current digital era. Rising instances of online harassment against women in Dravidian language communities have become a growing concern. The increase of anonymous accounts on social media platforms enables harmful behavior to spread, emphasizing the urgent need for robust detection systems to address this issue and protect vulnerable users. The research began with the implementation of a comprehensive preprocessing framework, incorporating thorough data cleaning, normalization, and tokenization processes. For feature engineering, we employed IndicBERT's embedding techniques to effectively capture essential linguistic patterns. While contemporary research typically gravitates towards transformer architectures for their perceived advantages, our findings highlight the efficiency of well-optimized IndicBERT-based methods.

The systematic evaluations of traditional machine learning algorithms were conducted, including Logistic Regression, SVM, and Random Forest classifiers (Priyadharshini et al., 2022). To provide a comprehensive analysis, we extended our investigation to include cutting-edge transformer-driven models such as the mBERT model, XLM-RoBERTa model, and IndicBERT model (Hariharan et al., 2024), each fine-tuned specifically for binary classification tasks. In recent times, transformer models like IndicBERT have demonstrated superior performance. Notably, in our case, the IndicBERT-based model achieved the highest accuracy, surpassing even mBERT. This highlights the effectiveness of language-specific pretraining in enhancing classification performance.

The major contributions of the paper are as follows:

- To predict abusive content against women, classical text encoding techniques are first used in the embedding layer and then Machine Learning (ML) algorithms are utilized.
- To explore transformer-based embedding model with Deep Learning (DL) model, we used mBERT, XLM-Roberta and IndicBERT for contextual feature extraction and then FFN is used for abusive content detection against women.
- To test the working of the proposed model, we utilized the benchmark dataset of abusive content against women.

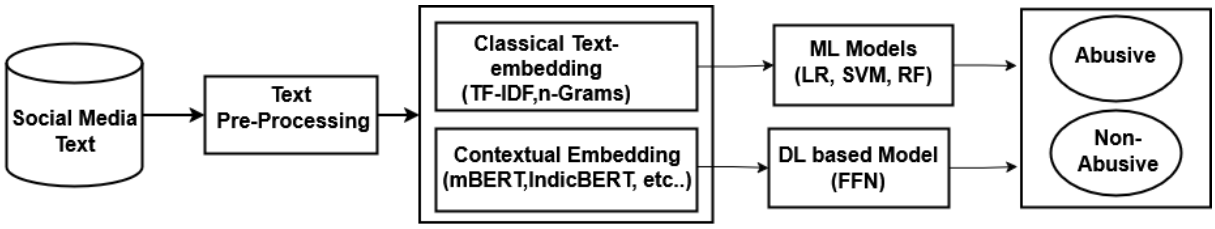


Figure 1: Architecture of the proposed model for ACD.

2 Literature Survey

Research in abusive language detection against women has shown varied approaches. Chakravarthi and Priyadharshini (Chakravarthi et al., 2023) found classical Machine Learning models like Logistic Regression outperforming deep learning on Fine-Grained Abusive Comment Detection (FGACD) due to limited data. Gupta and Roychowdhury (Gupta et al., 2022) improved performance using Term Frequency-Inverse Document Frequency (TF-IDF) for extracting features. Vegupatti et al. (Vegupatti et al., 2023) , Premjith et al. (Premjith et al., 2023) , and Hariharan et al (Hariharan and Anand Kumar, 2022) has focused on leveraging advanced BERT-based models like IndicBERT, MuRIL, and mBERT for multilingual content analysis. Their studies primarily target detecting abusive content against women and fake news across Indian languages, demonstrating the capability of deep learning methods to comprehend and classify text across diverse linguistic contexts.

3 Methodology

The architecture of the proposed model for abusive content detection against women is shown in Figure 1. Throughout the work, IndicBERT, a pretrained language model, is utilized in the embedding layer alongside TF-IDF. Notably, IndicBERT achieved the highest Macro F1-score, demonstrating its effectiveness in capturing linguistic patterns and enhancing classification performance.

3.1 Problem Definition

Abusive language detection against women is framed as a binary classification problem. Considering a dataset $C = \{c_1, c_2, \ldots, c_k\}$ consisting of k social media comments, each comment $c_i \in C$ is linked to a class label $y \in \{\text{non-abusive}(0), \text{abusive}(1)\}$. Each comment c_i contains p sentences $\{s_1, s_2, \ldots, s_p\}$, where each sentence $s_p \in c_i$ consists of q words $\{w_{j1}, w_{j2}, \ldots, w_{jq}\}$. A classification model $f : C \rightarrow y$ is defined and trained to predict the class label $y_j \in \{0, 1\}$, where $y_j = 0$ represents non-abusive content, and $y_j = 1$ represents abusive content.

3.2 Data Preprocessing

Data preprocessing prepares the dataset for training an Machine Learning model to detect abusive language in Tamil and Malayalam. The dataset is loaded, class labels are normalized, and missing values are replaced with empty strings. Text is cleaned by removing special characters, punctuation, and numbers. Removing noisy data by stabilizing the dataset for further usage.

Handling low-resource languages presents challenges due to limited annotated data and complex morphology. Key preprocessing includes Unicode normalization, subword tokenization, and noise filtering. Stopword removal eliminates Tamil and Malayalam stopwords, and English words are removed for consistency. The dataset is equalized across classes to prevent overfitting, ensuring a balanced dataset for effective abusive content detection.

3.3 Embedding Layer

The embedding layer converts text into numerical representations after removing Tamil and Malayalam stopwords to enhance relevance.

3.3.1 Traditional Text Encoding Techniques

TF-IDF is used where every word is given a weight based on its occurrence in a comment and its scarcity across the dataset (Shanmugavadivel et al., 2022). This method ensures the model prioritizing words that are most significant for distinguishing between abusive and non-abusive content for women.

3.3.2 Transformer-Based Embedding Techniques

The vectorized text data is divided into training and testing sets, enabling both model learning and performance evaluation. To improve the detection of abusive language targeting women, **pre-trained transformer-based embeddings** (such as BERT, XLM-R, or IndicBERT) are fine-tuned using a domain-specific dataset. This fine-tuning process involves updating model parameters through supervised learning, where abusive and non-abusive samples in Tamil and Malayalam for women help the model refine its contextual understanding.

Transformer models employ the **self-attention mechanism (SAM)** to identify relationships between words in a sequence, even over long distances. SAM assigns attention scores to determine the relative importance of different tokens, as described in Eq. (1):

$$\mathrm{A} = \text{softmax}\left(\frac{qk^\top}{\sqrt{d_k}}\right)v \tag{1}$$

where **A** denotes the attention matrix, **q** represents the query matrix, **k** is the key matrix, **v** corresponds to the value matrix, and $\mathbf{d_k}$ is the dimensionality of the key vectors.

Fine-tuning allows the transformer model to adapt to linguistic characteristics and abusive language patterns in Tamil and Malayalam. A classification layer, placed on top of the transformer, further enhances the model's ability to differentiate between abusive and non-abusive text based on Linguistic patterns. The model's performance is optimized by fine-tuning hyperparameters and evaluating key metrics such as precision, recall, and macro F1-score.

3.4 Classification Layer

After preprocessing and vectorization, the data is passed through classification models of ML and DL models to predict if a comment is abusive for women.

3.4.1 ML Models

Machine learning models being used are Logistic Regression, SVM, and Random Forest are used for classifying the text for abusive or non-abusive. The importance of classification is that it becomes easier to predict the input data.

3.4.2 DL Models

Deep learning models being used are multilingual Bert model, XLM-Roberta model and IndicBert model have been used. These models are used in embedding layer to efficiently represent data while capturing semantic relationships.

4 Experiment

This section provides the summary of the benchmark dataset utilized in this study, along with details of the experimental setup, dataset and performance metrics.

4.1 Experimental Setup

The experiment uses ML and DL models, implemented and tested in Jupyter Notebook, which integrates code, visuals, and text for streamlined computation, debugging, and visualization. TensorFlow, with its high-level API Keras, supports model implementation for architectures like mBERT, XLM-RoBERTa, and IndicBERT. The dataset is divided into 75% training, 25% testing, ensuring sufficient data for learning as well as performance evaluation.

Table 1: Summary of Datasets

Language	Class	Number of Samples
Tamil	Abusive	1366
	Non-abusive	1424
Malayalam	Abusive	1531
	Non-abusive	1402

4.2 Dataset

The dataset utilized in this study is well-defined and curated for accurate modeling and evaluation as shown in Table 1 (Premjith et al., 2023; Priyadharshini et al., 2022). It comprises YouTube comments in Tamil and Malayalam, annotated as Abusive or Non-Abusive content, specific to women as shown in Table 2. The data was sourced from publicly available comments and manually labeled based on linguistic and contextual cues.

Abusive comments contain derogatory, offensive, or misogynistic language, while non-abusive ones do not. The dataset also handles code-mixed text (e.g., Tanglish, Manglish). To prevent overfitting, we balanced the dataset by equalizing samples across both classes, ensuring unbiased learning. It was split into training, validation, and test-

Table 2: Example of Abusive and Non-Abusive Content in Tamil and Malayalam dataset

Type	Tamil	Malayalam
Abusive	ஆமா பா கட்டி வச்சி தோல உரிக்கணும்	നിനക്ക് വേണ്ടി വോട്ട് ചെയ്തവരെ തിരിഞ്ഞു കൊത്തുന്നു
Non-Abusive	எப்படி இருக்கிறீர்கள்?	സൂരജന്റെ കല്യാണം എപ്പോൾ നടക്കു അപ്പോൾ അറിയാം.....

ing sets for evaluation. Model performance was assessed using precision, recall, and macro F1-score to ensure reliability.

For reproducibility, we use datasets from prior research (Priyadharshini et al., 2022).

4.3 Performance metrics

Metrics on which the models have been evaluated are accuracy, precision, Recall, F1-Score and macro F1-Score from Eq - (2) to (6).

$$\mathrm{A} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

$$F1 = 2 \cdot \left(\frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \right) \tag{5}$$

$$\text{Macro-F1} = \frac{1}{N} \sum_{i=1}^{N} 2 \cdot \left(\frac{\mathrm{P}_i \cdot \mathrm{R}_i}{\mathrm{P}_i + \mathrm{R}_i} \right) \tag{6}$$

Where, **A**: Accuracy, **P**: Precision, **R**: Recall, **TP**: True Positive, **TN**: True Negative, **FP**: False Positive, **FN**: False Negative, **N**: Number of classes, $\mathbf{P_i}$: precision for class i, and $\mathbf{R_i}$: recall for class i.

5 Results and Analysis

This research investigates the performance of different models on two practical datasets in Tamil and Malayalam, focusing on key evaluation metrics like **Accuracy**, **Precision**, **Recall**, and **macro F1-score**. As illustrated, Table 3 shows the results of various models in terms of Accuracy, Precision, Recall, F1-Score and Macro F1-Score. The evaluation involved comparing traditional ML models, multilingual transformer-driven DL models, and cutting-edge techniques through two distinct experimental setups, aiming to identify the most effective model for detecting abusive content against women in these languages.

The findings reveal that the IndicBERT model outperforms other models across both datasets. This advantage stems from IndicBERT's ability to leverage pre-trained contextual embeddings, making it well-suited for handling code-mixed text and morphologically rich languages like Tamil and Malayalam. By capturing complex linguistic patterns without relying on manual feature engineering, IndicBERT effectively enhances performance on low-resource abusive content detection tasks. These results emphasize the importance of transformer-based models in such challenging scenarios.

5.1 Performance Comparison of Traditional and Transformer-Based Models for Tamil Text Classification

The evaluation of traditional embeddings like **TF-IDF** (Shanmugavadivel et al., 2022) and transformer-based models such as **mBERT**, **XLM-RoBERTa**, and **IndicBERT** (Reshma et al., 2023) demonstrated the superiority of transformer models in detecting abusive content as illustrated in Figure 2, particularly targeting women in Tamil text. Among these, **IndicBERT** emerged as the most effective model, achieving a **Macro F1-score of 0.750** for Tamil dataset. Its multilingual pretraining enabled it to grasp complex linguistic structures and contextual variations in Tamil, making it particularly well-suited for abusive content detection in these low-resource languages.

Traditional models like Random Forest, SVM and Logistic Regression struggled to identify im-

Table 3: Performance Comparison of the Proposed Model on Tamil and Malayalam Datasets

Model		Tamil					Malayalam				
Embedding	**Classifier**	**A**	**P**	**R**	**F1**	**Macro F1**	**A**	**P**	**R**	**F1**	**Macro F1**
TF-IDF	LR	0.669	0.670	0.668	0.669	0.669	0.661	0.671	0.648	0.652	0.659
TF-IDF	SVM	0.634	0.633	0.633	0.650	0.633	0.659	0.659	0.659	0.650	0.659
TF-IDF	RF	0.631	0.631	0.630	0.636	0.631	0.623	0.621	0.618	0.600	0.619
mBERT	FFN	0.748	0.748	0.746	0.746	0.745	0.685	0.680	0.680	0.682	0.685
XLM-R	FFN	0.730	0.730	0.730	0.733	0.735	0.700	0.705	0.700	0.703	0.705
IndicBERT	FFN	**0.750**	**0.750**	**0.750**	**0.750**	**0.750**	**0.720**	**0.725**	**0.720**	**0.720**	**0.720**

plicit abuse due to their reliance on manual features. In contrast, **IndicBERT** leveraged deep contextual understanding to detect subtle, context-dependent toxicity, outperforming them in Tamil. This highlights the advantage of transformers in low-resource language abuse detection.

5.2 Performance Comparison of Traditional and Transformer-Based Models for Malayalam Abusive Content Detection

For the Malayalam dataset, a comparative analysis of traditional embeddings (**TF-IDF** and **Hugging Face**) and transformer-based models (**mBERT**, **IndicBERT**, and **XLM-RoBERTa**) (Reshma et al., 2023) demonstrated the effectiveness of transformer-driven approaches in detecting abusive content as illustrated in Figure 2 particularly targeting women. Among these, **IndicBERT** emerged as the top performer, achieving a **Macro F1-score of 0.720** for Malayalam. Its multilingual pretraining allowed it to grasp the contextual intricacies of abusive language, making it highly effective in distinguishing both implicit and explicit forms of abuse.

Traditional models like Random Forest and SVM performed well in classification but struggled with the complexity of abusive language due to their reliance on manual features. In contrast, **IndicBERT** leveraged deep contextual embeddings to detect subtle abuse, excelling in identifying gender-targeted toxicity. This underscores the importance of transformers in addressing online abuse in Malayalam, where traditional methods often fall short.

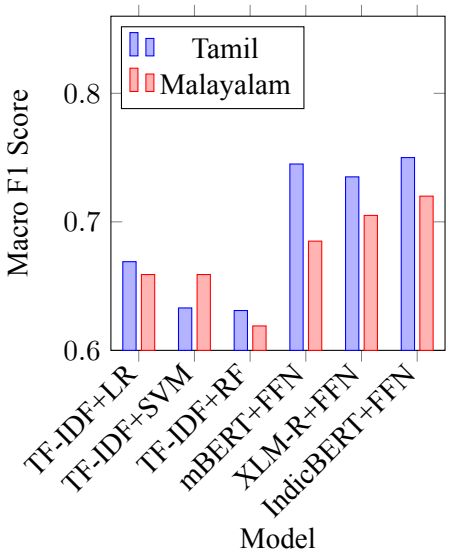


Figure 2: Comparison of Models on Macro F1-Score for Tamil and Malayalam Datasets

6 Conclusion and Future work

This study provides a comprehensive evaluation of Machine Learning and Deep Learning models for abusive content detection in Tamil and Malayalam. Contrary to previous trends favoring traditional machine learning models, transformer-based approaches, particularly **IndicBERT**, demonstrated superior performance in detecting gender-targeted toxicity. **IndicBERT** achieved the highest **Macro F1-score** of **0.750** for Tamil and **0.720** for Malayalam, effectively capturing both explicit and implicit forms of abuse. In contrast, traditional models like Random Forest, SVM, and Logistic Regression, which rely on manually engineered features, struggled to handle the complex and evolving nature of abusive language.

Despite these advancements, challenges remain in understanding model errors and improving robustness. Future research should focus on error analysis, real-world testing, and refining preprocessing to enhance transformer models for low-resource languages like Tamil and Malayalam. Additionally, incorporating external linguistic resources, improving contextual understanding, and mitigating biases within datasets could further enhance model effectiveness. Addressing domain-specific variations and handling code-mixed language more efficiently are also crucial for making these models more adaptable to real-world applications.

References

B. R. Chakravarthi, R. Priyadharshini, S. Banerjee, M. B. Jagadeeshan, P. K. Kumaresan, R. Ponnusamy, S. Benhur, and J. P. McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

V. Gupta, S. Roychowdhury, M. Das, S. Banerjee, P. Saha, B. Mathew, and A. Mukherjee. 2022. Multilingual abusive comment detection at scale for indic languages. In *Advances in Neural Information Processing Systems*, volume 35, pages 26176–26191.

R. I. L. Hariharan and M. Anand Kumar. 2022. Impact of transformers on multilingual fake news detection for tamil and malayalam. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 196–208. Springer.

R. L. Hariharan, M. Jinkathoti, P. S. P. Kumar, and M. A. Kumar. 2024. Fake news detection in telugu language using transformers models. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE.

B. Premjith, V. Sowmya, B. R. Chakravarthi, R. Natarajan, K. Nandhini, A. Murugappan, B. Bharathi, M. Kaushik, and P. Sn. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79.

R. Priyadharshini, B. R. Chakravarthi, S. Cn, T. Durairaj, M. Subramanian, K. Shanmugavadivel, S. Hegde, and P. Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the shared task on abusive tamil and malayalam text targeting women on social media: Dravidianlangtech@naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

S. Reshma, B. Raghavan, and S. J. Nirmala. 2023. Mitigating abusive comment detection in tamil text: A data augmentation approach with transformer model. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 460–465.

K. Shanmugavadivel, S. U. Hegde, and P. K. Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022. *DravidianLangTech*, page 292.

M. Vegupatti, P. K. Kumaresan, S. Valli, K. K. Ponnusamy, R. Priyadharshini, and S. Thavaresan. 2023. Abusive social media comments detection for tamil and telugu. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 174–187. Springer.

The_Deathly_Hallows@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages

Kogilavani Shanmugavadivel[1] **, Malliga Subramanian**[2]**,**
Vasantharan K[1]**,Prethish G A**[1]**,Santhosh S**[3]
[13]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{vasantharank.ncc, prethish0409, santhosh42169}@gmail.com

Abstract

The DravidianLangTech@NAACL 2025 shared task focused on multimodal hate speech detection in Tamil, Telugu, and Malayalam using social media text and audio. Our approach integrated advanced preprocessing, feature extraction, and deep learning models. For text, preprocessing steps included normalization, tokenization, stopword removal, and data augmentation. Feature extraction was performed using TF-IDF, Count Vectorizer, BERT-base-multilingual-cased, XLM-Roberta-Base, and XLM-Roberta-Large, with the latter achieving the best performance. The models attained training accuracies of 83% (Tamil), 88% (Telugu), and 85% (Malayalam). For audio, Mel Frequency Cepstral Coefficients (MFCCs) were extracted and enhanced with augmentation techniques such as noise addition, time-stretching, and pitch-shifting. A CNN-based model achieved training accuracies of 88% (Tamil), 88% (Telugu), and 93% (Malayalam). Macro F1 scores ranked Tamil 3rd (0.6438), Telugu 15th (0.1559), and Malayalam 12th (0.3016). Our study highlights the effectiveness of text-audio fusion in hate speech detection and underscores the importance of preprocessing, multimodal techniques, and feature augmentation in addressing hate speech on social media.

1 Introduction

Social media has revolutionized communication, allowing global information sharing. However, hate speech targeting groups based on race, religion, gender, or political views has also surged, presenting serious societal issues. Detecting hate speech, especially in low resource languages, is crucial for machine learning and NLP. Dravidian languages (Tamil, Telugu, and Malayalam) pose significant challenges due to their linguistic complexity and limited computational resources. Furthermore, these languages are often code-mixed with English, complicating NLP tasks. Effective hate speech detection requires multimodal models, as social media content is often a combination of text, audio, and images.

The DravidianLangTech@NAACL 2025 Multimodal Hate Speech Detection Shared Task addressed these challenges by providing datasets in Malayalam, Tamil, and Telugu, categorizing hate speech into Gender, Political, Religious, Personal Defamation, and Non-Hate Lal G et al. 2025.

We used extensive preprocessing, such as IndicNormalizerFactory for normalization, 'indic_tokenize.trivial_tokenize' for tokenization, stopword removal, and nlpaug for data augmentation. For audio, we extracted Mel-Frequency Cepstral Coefficients (MFCCs) using librosa and applied techniques like noise addition, time-stretching, and pitch-shifting. Feature extraction was done using TF-IDF, Count Vectorization, and XLM-Roberta-Large embeddings, while a CNN-based model analyzed MFCCs for audio classification.

Our models ranked third in Tamil, twelfth in Malayalam, and fifteenth in Telugu in the shared task. Despite challenges like class imbalances and noisy annotations, our results demonstrate that advanced preprocessing, feature extraction, and deep learning can effectively tackle these multimodal issues. This paper discusses our methods, challenges, and lessons learned in improving hate speech detection for low-resource languages.

2 Literature Review

Sreelakshmi et al. 2024 used machine learning classifiers and multilingual embeddings for hate speech detection in CodeMix Dravidian languages. MuRIL performed best for Malayalam, and a cost-sensitive strategy addressed class imbalance. A new annotated Malayalam-English dataset was introduced.

Premjith et al. 2024a highlighted challenges in Telugu CodeMix text and the need for hate speech detection. MPNet and CNNs were used in the HOLD-Telugu shared task, utilizing macro F1-scores for binary classification. Premjith et al. 2024b showed monolingual models' limitations for Tamil and Malayalam and emphasized multilingual BERT and multimodal analysis (text, audio, video) for better performance. Chakravarthi et al. 2023 proposed a fusion model combining MPNet and CNN for detecting offensive language in CodeMix Dravidian languages, achieving high F1-scores (Tamil: 0.85, Malayalam: 0.98, Kannada: 0.76). Imbwaga et al. 2024 explored hate speech detection in English and Kiswahili audio. Spectral, temporal, and prosodic features were extracted, with XGBoost excelling in Kiswahili and Random Forest in English, highlighting language-specific classifier importance. Premjith et al. 2023 discussed a Dravidian Languages Workshop task on multimodal abusive language detection and sentiment analysis. Sixty teams participated, using macro F1-scores for evaluation. Barman and Das 2023 addressed abusive language detection in Tamil and Malayalam, achieving an F1-score of 0.5786 using mBERT, ViT, and MFCC.

An et al. 2024 introduced two explainable audio hate speech detection methods: End-to-End (E2E) and a cascade approach. E2E performed better, with frame-level justifications improving accuracy. Koreddi et al. 2024 presented an AI system for detecting objectionable content across text, audio, and visual media, integrating speech recognition, OCR, NLP, and Google Translator. BERT was used for text analysis, enhancing online content safety. Narula and Chaudhary 2024 studied hate speech detection challenges in Hindi, focusing on dialects, code-switching, and Romanized Hindi. Advanced machine learning was suggested to combat misinformation and societal unrest.

3 Task Description

The Multimodal Hate Speech Detection Shared Task at DravidianLangTech@NAACL 2025 challenges researchers to detect hate speech in Tamil, Malayalam, and Telugu using multimodal social media data. Participants receive training data with text and speech components to classify hate speech into gender, political, religious, and personal defamation categories. Models are evaluated using the macro-F1 score.

4 Dataset Description

Anilkumar et al. 2024 present a dataset comprising YouTube videos in Tamil, Malayalam, and Telugu, incorporating text and audio features. Hate speech is classified into gender (G), political (P), religious (R), and personal defamation (C) categories. Text preprocessing includes tokenization, stopword removal, and augmentation, while audio features are extracted using MFCC with enhancements like noising, time-stretching, and pitch-shifting.

Language	Non-Hate	Hate (C,G,P,R)
Malayalam	406	477
Tamil	287	227
Telugu	198	358

Table 1: Text Data Distribution

Language	Non-Hate	Hate (C,G,P,R)
Malayalam	406	477
Tamil	287	222
Telugu	198	353

Table 2: Audio Data Distribution

5 Methodology

The multimodal hate speech detection approach categorizes content as hate or non-hate speech, with subclasses for gender, political, religious, and personal defamation. Text and audio data from YouTube videos undergo preprocessing, including tokenization, stopword removal, augmentation, and MFCC-based audio feature extraction with noise addition and pitch/time-stretching. Text features are derived using TF-IDF, Count Vectorizer, and transformer embeddings (e.g., XLM-Roberta), while MFCCs are used for audio. A fully connected model classifies text, while CNN handles audio. Optimization techniques such as dropout, and early stopping enhance performance. Models are evaluated using the macro F1 score, addressing linguistic and multimodal challenges while ensuring scalability.

5.1 Data Preprocessing

Preprocessing enhances model performance by cleaning, normalizing, and augmenting text and audio data, ensuring consistency and meaningful feature extraction.

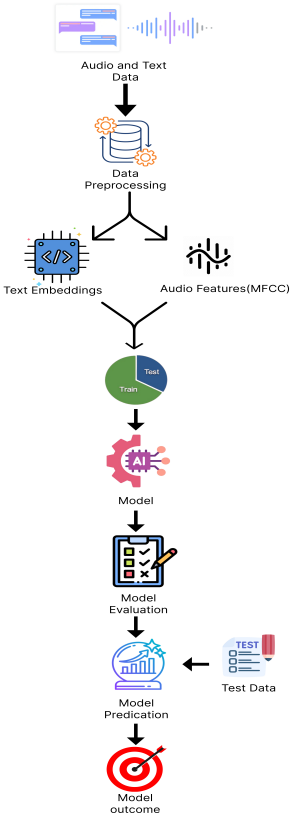


Figure 1: Proposed Model Workflow

5.1.1 Text Preprocessing and Feature Extraction

Tamil, Telugu, and Malayalam text is normalized using IndicNormalizerFactory to handle script variations. Tokenization is performed with 'indic_tokenize.trivial_tokenize', and stopwords are removed using custom lists (Malayalam) and advertools (Tamil, Telugu). nlpaug is used for data augmentation (synonym replacement, word insertion). For feature extraction, TF-IDF, Count Vectorizer, and transformer embeddings (XLM-Roberta-Base, XLM-Roberta-Large, BERT Multilingual Cased) are used. XLM-Roberta-Large provides the best accuracy due to its strong multilingual capabilities.

5.1.2 Audio Preprocessing and Feature Extraction

Audio features are extracted using MFCCs, which mimic human auditory perception. Augmentation techniques (noise addition, pitch/time-stretching) improve model robustness. These processed features ensure stability under varying conditions. These processed features were essential in training effective multimodal models.

5.2 Model Architecture

The multimodal system classifies hate speech using separate architectures for text and audio.

5.2.1 Text Classification Model

XLM-Roberta-Large and BERT-Multilingual Cased embeddings capture contextual features. The model includes dense layers with ReLU, batch normalization, dropout, and a softmax output layer for classification (Gender, Political, Religious, Personal Defamation Hate, Non-Hate). Optimized with Adam, categorical cross-entropy, and evaluated using macro-F1 score.

5.2.2 Audio Classification Model

A CNN-based model processes 2D MFCC features from speech data. Convolutional layers with increasing filters extract hierarchical patterns, while dropout after pooling prevents overfitting. A dense layer refines features before the softmax output. Training is optimized with ReduceLROnPlateau, EarlyStopping, ModelCheckpoint.

Together, these models form a robust multimodal hate speech detection system, excelling in both text and audio classification.

6 Limitations

Although this study of demonstrates strong performance, certain limitations remain. The primary challenge is the limited dataset size, which affects the model's ability to generalize across diverse hate speech patterns. Additionally, class imbalance in the dataset impacts the fair representation of all categories, leading to biased classification. While data augmentation helps improve model robustness, it cannot fully compensate for the lack of real-world diversity in the dataset. Future work can focus on expanding the dataset and implementing more effective balancing techniques to enhance performance.

7 Performance Evaluation

The models were evaluated using classification metrics such as precision, recall, F1-score, and accuracy to assess their effectiveness in detecting hate and non-hate speech across languages. Precision measured the model's ability to minimize false positives, while recall indicated its capacity to detect relevant samples. The F1-score, a harmonic mean of precision and recall, was crucial for handling class imbalances. Accuracy represented the proportion of correctly predicted instances. The best per-

Models Used	Tamil	Telugu	Malayalam
BERT-base-multilingual-cased	67%	71%	73%
TF-IDF Vectorizer	55%	58%	58%
CountVectorizer	59%	60%	58%
XLM-Roberta-base	73%	71%	72%
XLM-Roberta-large	83%	88%	85%

Table 3: Performance of Different Models in Tamil, Telugu, and Malayalam (text)

Models Used	Tamil	Telugu	Malayalam
Without Preprocessing	64%	54%	85%
TTSAudio + Dynamic Normalizer	58%	54%	80%
Data Augmentation	85%	82%	90%
Data Augmentation with BatchNormalization	88%	88%	93%

Table 4: Performance of Different Models in Tamil, Telugu, and Malayalam (audio)

formance in text classification was achieved using XLM-Roberta-Large embeddings, while MFCCs and augmentation improved audio classification. These metrics provided a comprehensive evaluation of the models across multiple modalities and languages.

8 Conclusion

This study aimed to develop a robust multimodal hate speech detection system for Malayalam, Tamil, and Telugu. Using textual data with "XLM-Roberta-Large" embeddings and audio data with "MFCCs" and data augmentation, the system categorized hate speech into gender, political, religious, and personal defamation. The models achieved third place in the shared challenge, with Tamil scoring 0.6438, and Telugu and Malayalam scoring 0.1559 and 0.3016, respectively. Despite challenges like unbalanced datasets and complex cultural contexts, this work demonstrates how multimodal techniques can address hate speech detection in under-represented languages. Future work could explore larger datasets, additional modalities, specialized architectures for multimodal tasks, and better domain adaptation strategies to further enhance performance. This study provides a foundation for ensuring linguistic inclusion, improving online safety, and enabling fully automated hate speech detection. The source code for our approach is available at https://github.com/vasantharan/Multimodal_Hate_Speech_Detection_in_Dravidian_languages.

References

Jinmyeong An, Wonjun Lee, Yejin Jeon, Jungseul Ok, Yunsu Kim, and Gary Geunbae Lee. 2024. An investigation into explainable audio hate speech detection. *arXiv preprint arXiv:2408.06065*.

Abhishek Anilkumar, Jyothish Lal G, B Premjith, and Bharathi Raja Chakravarthi. 2024. Dravlangguard: A multimodal approach for hate speech detection in dravidian social media. In *Speech and Language Technologies for Low-Resource Languages (SPELLL)*, Communications in Computer and Information Science.

Shubhankar Barman and Mithun Das. 2023. hate-alert@ dravidianlangtech: Multimodal abusive language detection and sentiment analysis in dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 217–224.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

Joan L Imbwaga, Nagatatna B Chittaragi, and Shashidhar G Koolagudi. 2024. Automatic hate speech detection in audio using machine learning algorithms. *International Journal of Speech Technology*, 27(2):447–469.

Venkatesh Koreddi, Nalluri Manisha, Shaik Mohammad Kaif Mohammad Kaif, and Yeligeti Tejaswa Sai Kumar. 2024. Multilingual ai system for detecting offensive content across text, audio, and visual media. *Engineering Research Express*.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Nataraj an, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech

Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Rachna Narula and Poonam Chaudhary. 2024. A comprehensive review on detection of hate speech for multi-lingual data. *Social Network Analysis and Mining*, 14(1):1–35.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

B Premjith, V Sowmya, Bharathi Raja Chakravarthi, Rajeswari Natarajan, K Nandhini, Abirami Murugappan, B Bharathi, M Kaushik, Prasanth Sn, et al. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

SSN_IT_NLP@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media

Maria Nancy C[1], **Radha N** [2], **Swathika R**[3]
[1]Annai Veilankanni's College of Engineering, Nedungundram , India
[2,3] Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, India
nancycse13@gmail.com[1]
radhan@ssn.edu.in[2]
swathikar@ssn.edu.in[3]

Abstract

The proliferation of social media platforms has led to a rise in online abuse, particularly against marginalized groups such as women. This study focuses on the classification of abusive comments in Tamil and Malayalam, two Dravidian languages widely spoken in South India. Leveraging a multilingual BERT model, this paper provides an effective approach for detecting and categorizing abusive and non-abusive text. Using labeled datasets comprising social media comments, our model demonstrates its ability to identify targeted abuse with promising accuracy. This paper outlines the dataset preparation, model architecture, training methodology, and the evaluation of results, providing a foundation for combating online abuse in low-resource languages. This methodology uniquely integrates multilingual BERT and weighted loss functions to address class imbalance, paving the way for effective abuse detection in other underrepresented languages. The BERT model achieved an F1-score of 0.6519 for Tamil and 0.6601 for Malayalam. The code for this work is available on github Abusive-Text-targeting-women.

1 Introduction

Social media platforms have grown to be an important online forum for entertainment, communication, and information exchange in recent years. Despite their advantages, these platforms are increasingly misused to target women with derogatory language. Due to cultural biases and gender inequality, women are frequently the victim of cruel and disparaging remarks that aim to denigrate, harass, or threaten them. Women may face significant psychological, social, and professional repercussions from this form of online abuse, a distinct type of cyberbullying that necessitates appropriate intervention. By identifying offensive language directed at women in comments, we address this issue by concentrating on online content management. In order to complete this assignment, we used targeted searches to scrape YouTube comments on sensitive and contentious subjects where gender-based abuse is common(Rajiakodi et al., 2025).These queries targeted explicit abuse, implicit bias, stereotypes, and coded language. This task's objective is to determine whether or not a certain comment contains abusive language. Text in the low-resource South Indian languages of Tamil and Malayalam is included in the dataset.

2 Related Work

The fast growth of social media has amplified the existence of online abuse, which targets more women as they mainly represent the underprivileged group and suffer more and in different ways (Chakravarthi et al., 2023). However, the two Dravidian languages are still underrepresenting themselves in the domain of computer linguistics, particularly when it comes to identifying abusive content (Mohan et al., 2023). This paper brings a highly effective approach to comment classification, abused as well as non-abused, for those languages with the use of a multilingual BERT model. This is an approach that does well on using labeled datasets and weighted loss functions for handling class imbalance (Shanmugavadivel et al., 2022). This paper tries to improve the detection of abuse in low-resource languages, giving important insights that will open avenues for further research on fighting online abuse in different linguistic domains (Rajalakshmi et al., 2023). This study is different from the previous ones because it is based on Tamil and Malayalam, which are languages with complicated patterns, hence making it hard to identify misuse (Subramanian et al., 2023). Multilingual BERT ensures that the model can efficiently handle code-mixed text and regional linguistic peculiarities (Ponnusamy et al., 2024). In addition, weighted loss functions and thorough preprocess-

ing stages enhance the robustness of the proposed method (Shanmugavadivel et al., 2023).Limitations on the use of digital spaces include harassment against women and other disadvantage groups online. In the development of abuse classifiers for the ICON2023 Gendered Abuse Detection challenge, annotated Twitter datasets in English, Hindi, and Tamil were available. Combining two Ensemble Approach (EA), namely CNN and BiLSTM modeling contextual dependencies while CNN captures abusive language characteristics, an ensemble model is presented by the CNLP-NITS-PP team (Vetagiri et al., 2024). Comments on social media can shift the political and corporate climate overnight, affecting people and civilizations in ways that are impossible to ignore. But the same media also make it easier to launch targeted attacks on specific people or organizations. To identify hope speech and harmful remarks categorized as xenophobia, transphobia, homophobia, misogyny, misandry, and counter-speech, a shared job was implemented (Priyadharshini et al., 2022). Participants used a variety of Deep Learning (DL) and machine learning models on datasets that were either entirely Tamil or included a mix of Tamil and English codes. They then presented their findings and insights into how they used the datasets. The prevalence of social media calls for increased efforts toward the identification and classification of abusive remarks (Priyadharshini et al., 2023). In this process, there is an increasing need for Tamil and other low-resource Indic languages, which happen to provide greater obstacles (Reshma et al., 2023). The paper makes use of data augmentation methods like lexical substitution and back-translation along with multilingual transformer-based models for categorizing offensive comments posted on YouTube in Tamil. The Multilingual Representations for Indian Languages (MURIL) transformer model had the best performance with a 15-point improvement in macro F1-scores compared to baselines. This article discusses methods for efficient preparation of Tamil text for abuse detection.

3 Dataset Description

To classify abusive and non-abusive comments in Dravidian languages—Tamil and Malayalam—we used datasets consisting of development, training, and test data. Malayalam has 629 samples (Abusive-304, non-abusive-325); Tamil has 598 samples (Abusive-278, non-abusive-320); Two classifications comprise each datasets: Abusive and Non-Abusive. For every language, the test data consist of 558 Tamil and 629 Malayalam samples.

Language	Development Data	Training Data	Test Data
Tamil	598	2700	558
Malayalam	629	2933	629

Table 1: Details of the dataset used to classify abusive and non-abusive comments in Dravidian languages—Tamil and Malayalam.

4 Methodology

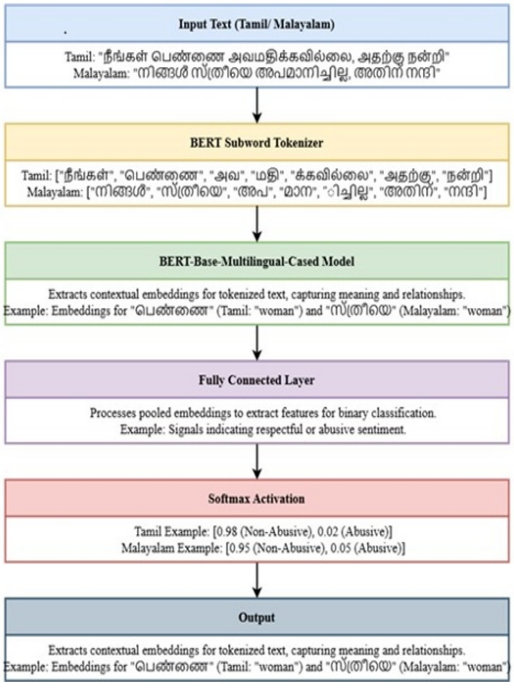


Figure 1: BERT Model Architecture

Figure 1 illustrates the model architecture that employs BERT for identifying offensive comments. Contextual embeddings are extracted from Tamil or Malayalam input text using BERT Subword Tokenizer for tokenization, and then it is processed by the BERT-base-multilingual-cased model. A Fully Connected Layer is applied for binary classification using these embeddings and a Softmax Activation function. The output layer categorizes the input as abusive or non-abusive, an efficient way of dealing with language variation and complexity.

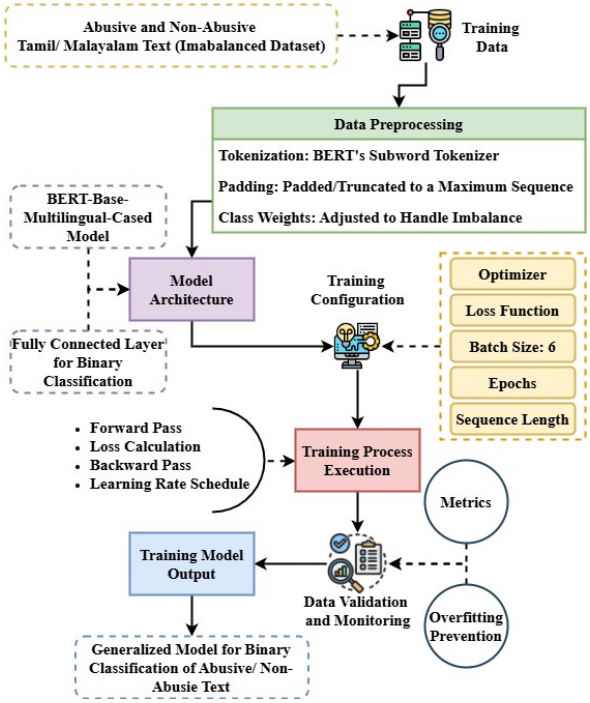


Figure 2: Training process of BERT-Base-Multilingual-Cased Model

Figure 2 presents a workflow for abusive and non-abusive text classification in Tamil and Malayalam using a BERT-based model. The process begins with training data, which consists of an imbalanced dataset of abusive and non-abusive text. In the data pre-processing step, the BERT sub-word tokenizer is used for tokenization, while the sequences are padded or truncated to a fixed length. Additionally, class weights are adjusted to handle the imbalance of the data set. The model architecture is based on the BERT-Base Multilingual Cased Model, enhanced with a fully connected layer for binary classification. The training configuration specifies key hyper parameters such as the optimizer, loss function, batch size (6), number of epochs, and sequence length. During training execution, the model undergoes a forward pass, loss calculation, backward pass, and learning rate scheduling to optimize performance. Once trained, the model outputs a generalized classifier capable of distinguishing abusive and non-abusive text. Performance is assessed using evaluation metrics, and data validation and monitoring ensure model reliability. In addition, overfitting prevention techniques are applied to improve generalization. This structured approach uses deep learning and transfer learning techniques to improve abusive speech detection in low-resource Dravidian languages.

$$Xs*Ba[ki-ah] \rightarrow Ls[v-zw'']+Va[\partial\alpha-aqw''] \quad (1)$$

The equation represents the transformations of input token embeddings through the self-attention and optimization mechanisms of the BERT-Base Multilingual Cased model for abusive and non-abusive text classification. Here, Xs denotes the token embeddings, which are numerical vector representations of words in the input text. $Ba[ki-ah]$ refers to the attention-weighted representation of these token embeddings, capturing contextual relationships between words. The transformation yields $Ls[v-zw'']$, an intermediate feature representation obtained from the hidden layers, where zw'' represents refinements made by the self-attention mechanism. Additionally, $Va[\partial\alpha-aqw'']$ describes the gradient-based optimization process during model training, where $\partial\alpha$ represents parameter updates in backpropagation, and aqw'' indicates higher-order interactions in contextual learning. Together, these components refine word representations, ensuring that abusive language patterns are effectively captured while maintaining contextual integrity. This transformation ultimately enables the classification layer to accurately differentiate between abusive and non-abusive text.

5 Result and Discussion

The results indicate that the proposed BERT-based model outperforms existing methods across key evaluation metrics, making it highly effective for detecting abusive comments in Tamil and Malayalam.

5.1 Evaluation Metrics for Tamil

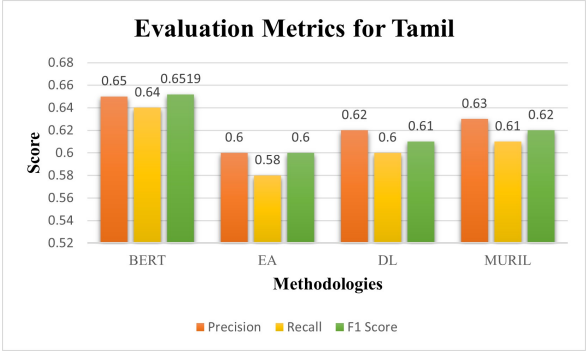


Figure 3: Evaluation Metrics for Tamil

The performance of BERT for the classification of abusive comments in Tamil achieved an F1 score of 0.6519, outperforming related methods such as EA (0.60), DL (0.61) and MURIL (0.62). This demonstrates BERT's ability to handle linguistic complexity and imbalanced datasets better. Preprocessing techniques, including script normalization and handling code-mixed text, significantly enhanced the classification quality. Challenges in capturing contextual nuances, such as sarcasm and implicit abuse, remain. Future work will explore advanced embedding techniques for better results.

$$Ds[\{li{-}an''\}] \rightarrow Ls[as{-}naq'']{+}Va[ds{-}iuwq''] \quad (2)$$

The equation describes the transformation of input data through the self-attention and optimization mechanisms in the BERT-Base Multilingual Cased model for abusive and non-abusive text classification. Here, $Ds[\{li{-}an''\}]$ represents the processed input embeddings, where li and an'' refer to specific token-level features, possibly modified by language embeddings or attention mechanisms. The right-hand side consists of two key components: $Ls[as - naq'']$, which denotes an intermediate feature representation extracted from the hidden layers, capturing the contextual relationships between words, and $Va[ds - iuwq'']$, which represents the gradient-based optimization process used for model fine-tuning. Here, ds may correspond to weight adjustments, while $iuwq''$ indicates advanced contextual refinements applied through backpropagation. This transformation ultimately helps the model better distinguish abusive text from non-abusive text by refining token relationships and optimizing the learned embeddings for classification.

5.2 Evaluation Metrics for Malayalam

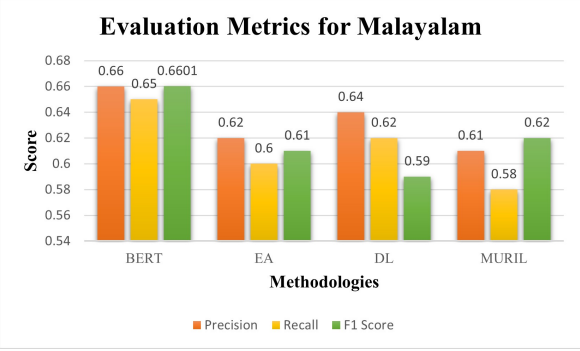


Figure 4: Evaluation Metrics for Malayalam

Although the F1-scores for EA, DL, and MURIL were all 0.66 for abusive comment classification in the Malayalam dataset, BERT achieved the highest F1-score of 0.6601. Using class-weighted loss functions that along with strong preprocessing greatly enhanced the performance of the model, therefore addressing the class imbalance and language variation. Although it beats other methods, the misclassification of more delicate cases of abusive language calls for improvement even if it exceeds other algorithms. Contextual embeddings and explainable artificial intelligence techniques help to identify abusive content in Malayalam even more.

$$p_{fvd}[k{-}anw''] \rightarrow Dsp[v{-}znq'']{+}Va[s{-}e6v''] \quad (3)$$

The equation represents a transformation in the BERT-Base Multilingual Cased model for abusive and non-abusive text classification, illustrating how token embeddings evolve through different layers of the model. Here, $p_{fvd}[k - anw'']$ denotes the initial token representation, where k represents token indices, and anw'' signifies preprocessed features, possibly modified by positional encodings and attention mechanisms. The right-hand side consists of two major components: $Dsp[v - znq'']$, which refers to an intermediate feature representation extracted from hidden layers after applying the self-attention mechanism, and $Va[s - e6v'']$, which describes the gradient-based optimization process in the fine-tuning stage of BERT. Here, s represents model parameters, while $e6v''$ indicates deeper refinements applied through backpropagation. Together, these transformations allow the model to accurately learn contextual relationships between words, thereby improving its ability to classify text as abusive or non-abusive.

6 Limitations

Only a limited sample of training data is used to train the model. Misclassification may result from difficulties identifying slang, sarcasm, and contextual meaning. Furthermore, the results of the study might not be entirely indicative of long-term trends or generalizable.

7 Conclusion

Using the BERT-Base Multilingual Cased model, this study demonstrates an efficient method for classifying abusive comments in Tamil and Malayalam.

Multilingual embeddings, combined with sophisticated preprocessing techniques and class-weighted loss functions, successfully addressed the challenges of language variation and class imbalance. The model achieved F1-scores of 0.6519 for Tamil and 0.6601 for Malayalam, outperforming the baseline models EA, DL, and MURIL. Significant efficiency improvements were achieved through preprocessing techniques such as script normalization and handling code-mixed text. Despite these advancements, challenges persist, particularly in the misclassifications of subtle expressions such as sarcasm and implicit abuse. Future work includes expanding the dataset with additional languages and diverse sources, developing multilingual embeddings tailored for low-resource languages, and incorporating explainable artificial intelligence to enhance interpretability. These advancements aim to improve abusive content detection systems, thereby enhancing online safety for vulnerable individuals.

References

B. R. Chakravarthi, R. Priyadharshini, S. Banerjee, M. B. Jagadeeshan, P. K. Kumaresan, R. Ponnusamy, and J. P. McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

J. Mohan, S. R. Mekapati, and B. R. Chakravarthi. 2023. A multimodal approach for hate and offensive content detection in Tamil: From corpus creation to model development. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

R. Ponnusamy, K. Pannerselvam, R. Saranya, P. K. Kumaresan, S. Thavareesan, S. Bhuvaneswari, and B. R. Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Siddhanth Shanmugavadivel, Kogilavani U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*.

R. Rajalakshmi, S. Selvaraj, and P. Vasudevan. 2023. Hottest: Hate and offensive content identification in Tamil using transformers and enhanced stemming. *Computer Speech & Language*, 78:101464.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

S. Reshma, B. Raghavan, and S. J. Nirmala. 2023. Mitigating abusive comment detection in Tamil text: A data augmentation approach with transformer model. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 460–465.

K. Shanmugavadivel, R. Chinnasamy, N. Subbarayan, A. Ganesan, D. Ravi, V. Palanikumar, and B. R. Chakravarthi. 2023. On finetuning adapter-based transformer models for classifying abusive social media Tamil comments.

K. Shanmugavadivel, S. U. Hegde, and P. K. Kumaresan. 2022. Overview of abusive comment detection in Tamil-acl 2022. In *DravidianLangTech 2022*, page 292.

M. Subramanian, K. Shanmugavadivel, N. Subbarayan, A. Ganesan, D. Ravi, V. Palanikumar, and B. R. Chakravarthi. 2023. On finetuning adapter-based transformer models for classifying abusive social media Tamil comments.

A. Vetagiri, G. Kalita, E. Halder, C. Taparia, P. Pakray, and R. Manna. 2024. Breaking the silence detecting and mitigating gendered abuse in Hindi, Tamil, and Indian English online spaces. *arXiv preprint arXiv:2404.02013*.

LinguAIsts@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media

Dhanyashree G[1], Kalpana K[2], Lekhashree A[3], Arivuchudar K[4], Arthi R[5], Bommineni Sahitya[6], Pavithra J[7], Sandra Johnson[8]
R.M.K. Engineering College, Tiruvallur, Tamilnadu, India
{dhan22012, kalp22020, lekh22026, ariv22002}.ad@rmkec.ac.in
{arth22004, bomm22009, pavi22039, hod}.ad@rmkec.ac.in

Abstract

Social networks are becoming crucial sites for communication and interaction, yet are increasingly being utilized to commit gender-based abuse, with horrific, harassing, and degrading comments targeted at women. This paper tries to solve the common issue of women being subjected to abusive language in two South Indian languages, Malayalam and Tamil. To find explicit abuse, implicit bias, preconceptions, and coded language, we were given a set of YouTube comments labeled Abusive and Non-Abusive. To solve this problem, we applied and compared different machine learning models, i.e., Support Vector Machines (SVM), Logistic Regression (LR) and Naive Bayes classifiers, to classify comments into the given categories. The models were trained and validated using the given dataset, achieving the best performance with an accuracy of 89.89% and a macro F1 score of 90% using the best-performing model. The proposed solutions aim to develop robust content moderation systems that can detect and prevent abusive language, ensuring safer online environments for women.

1 Introduction

Over the years, social networks have become an overwhelmingly popular channel for entertainment, communication, and distribution of information. But despite this advantage, it has also become a platform in which cyberbullying and harassment occur predominantly. Cyberbullying occurs in a major way among women, a reflection of deep-seated cultural prejudice or gender inequality, and it also often manifests itself in the form of nasty, vilifying, and threatening speech. Given the strong psychological, social, and professional consequences of this type of focused harassment, creating strong protections against such speech is absolutely necessary.

Malayalam and Tamil are two prominent languages used on social media platforms in South India. However, the resource-scarce nature adds to the challenges of effective content-moderation tools in these two languages. Inappropriate comments with low-resource languages usually include explicit language, implicit bias, stereotypes, and coded language that makes them more difficult to spot.

This research aligns with the shared task on the detection of abusive comments in Tamil and Telugu proposed by Priyadharshini et.al (2023). Their analysis provides a benchmark dataset and evaluations used to contribute to the advancement of abusive comment detection. Also, Priyadharshini et.al, in the DravidianLangTech@ACL (2022) workshop, discussed the impact of abusive language on social media and highlighted the challenges posed by code-mixed Tamil and English text. By incorporating these insights, our study contributes to ongoing efforts in low-resource language processing. It improves the accuracy of abuse detection systems and reinforces the need for multilingual AI-driven moderation tools.

The present research will identify gender-related abusive content in comments posted on the Malayalam and Tamil YouTube streams, with a focus on solving the concern. The goals of this project are to implement machine learning algorithms to classify comment categories as abusive and non-abusive using datasets that have received binary labels applied. The current data set used contains diverse abusive content, both explicit and implicit.We have used the support vector machine (SVM), logistic regression (LR) and Naive Bayes machine learning models to perform the classification task. For implementation, please refer to this GitHub repository(Dhanyashree-G).

2 Related Work

The Abusive Comment Detection in Tamil-ACL 2022 shared task consisted of an experiment by Balouchzahi et al.(2022) on detecting abusive comments in Tamil. To address challenges such as code mixing, context dependence, and data imbalance, their experiment considered abusive language in native Tamil script, as well as code-mixed Tamil texts. They proposed two models for the task: (i) a 1D Convolutional LSTM (1D Conv-LSTM) model and (ii) an n-gram Multilayer Perceptron model (n-gram MLP) utilizing char n-grams and performed well for Tamil mixed code with a weighted F1 score of 0.56. For detecting abusive content, the n-gram MLP model outperformed the 1D Conv-LSTM model. This paper illustrates how feature engineering and classical machine learning can be used to detect abusive content in low-resource, code-mixed languages.

The shared task of Chakravarthi et al. was discussed in his presentation shortly after the third publication. The (2021) project of offensive language identification in Tamil, Malayalam, and Kannada languages was conducted through the (2021), addressing the challenges of detecting abuse in under-resourced Dravidian languages. This task emphasized the importance of identifying offensive language in multilingual and code-mixed texts prevalent in user-generated content on social media platforms. The dataset for this task included six categories of annotations, such as Not offensive, offensive untargeted, and offensive. Participants used a wide variety of methodologies, including traditional machine learning algorithms, deep learning architectures, and transformer-based models. Pre-trained multilingual transformers such as mBERT, XLM-R, and IndicBERT have been carefully evaluated to classify offensive content. The model that performs best have achieved F1 scores of up to 0.97% for Malayalam and 0.78% for Tamil, highlighting the potential utility of transformer-based models in offensive language detection.

Rajalakshmi et al. solved the problem of detecting abusive comments in Tamil and Tamil-English datasets under the shared task DravidianLangTech@ACL 2022. The primary goal of the study was to detect abusive content categories such as homophobia, transphobia, xenophobia, and counter-speech that often form within the community. Three approaches were used by the authors: transformer-based models, deep learning (DL), and machine learning. Random Forest outperformed other algorithms with a weighted F1 score of 0.78% on its Tamil and English dataset. Pre-trained word embeddings with BiLSTM models performed better among deep learning models for Tamil data. mBERT was the best-performing transformer-based model with an F1 score of 0.70% for Tamil comments. Issues such as class imbalance and the dominance of code-mixed and code-switched data that make detection tasks more difficult were also addressed in the study. The authors published a paper using advanced techniques such as balanced class weights and fine-tuning transformer models to identify abusive content.

Pannerselvam et al. (2023) addressed the issue of identifying offensive remarks in code-mixed Tamil-English and Telugu-English text. They concentrated on developing a multiclass classification model that could distinguish between eight types of offensive remarks. The study used the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and two text representation techniques, Bag of Words (BOW) and Term Frequency Inverse Document Frequency (TF-IDF), to solve the problems caused by code-mixed data. Machine learning algorithms such as Support Vector Machine (SVM), Random Forest, and Logistic Regression were used to perform the categorization. It performed best among them with an F1 score of 0.99% TF-IDF representation and SMOTE-balanced data to achieve the highest performance. The study shows how mixing SMOTE and TF-IDF works well to handle unbalanced datasets and catch the subtle differences in mean speech across languages. Their method proved strong and looked good for real-world use in managing angry comments on online platforms. This was clear from how they came in ninth place in the shared task, even when dealing with issues like language gaps or changes in what people think is offensive.

The author, Zichao Li, has combined classes, adjusted course weights with respect to the reciprocal of log frequencies, and used focal loss to put more focus on the minority classes during training to tackle challenges such as class imbalance in the dataset. We applied additional adversarial training to improve the robustness and generalization ability of the model. This resulted in one of the top-performing systems with a weighted average F1 score of 0.75%, 0.94%, and 0.72%, individually placing it at fourth, third, and fourth in Tamil-English, Malayalam-English, and Kannada-English tasks, respectively. This work emphasizes the usefulness of transformer approaches for dealing with code-mixed texts in low-resourced languages such as Tamil, Malayalam, and Kannada through novel use of multilingual transformers, applicable preprocessing methods, and specialized loss functions.

3 Methodology

The dataset and the experiments we carried out for the study are described in depth in this section. The system architecture for classifying abusive comments into binary classes using machine learning (ML) methods, such as GridSearchCV, The general flow of the categorization process is shown in the figure 1.

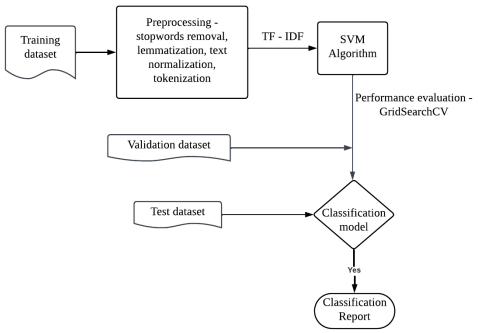


Figure 1: System Architecture for Detecting Abusive Comments Using ML Models.

3.1 Dataset

The dataset for this study consists of YouTube comments written in Tamil and Malayalam. Each language has its datasets divided into three subsets: train, validate, and test. The data sets are divided into two classes: Abusive and Non-Abusive. The detailed distribution of each dataset is showed in Table 1.

Language	Train	Validate	Test
Malayalam	2933	629	500
Tamil	2790	598	450

Table 1: The dataset distribution for Tamil and Malayalam, including the number of samples for each language.

The datasets were adapted and modified from publicly accessible datasets originally published as part of the DravidianLangTech@NAACL2025 program to suit the specific context of this study.

3.2 Proposed Solution

The proposed solution for detecting abusive and non-abusive comments in Tamil and Malayalam employs a combination of preprocessing techniques, feature engineering, and machine learning models to achieve precise and interpretable classification. Similarly, exploratory data analysis was integrated using WordCloud to visualize the most common abusive and non-abusive terms in the dataset, providing insight into language patterns.

3.3 Exploratory Data Analysis

Unbiased comments were generated in WordCloud for abusive and non-abusive comments compared to standard datasets. These visualizations highlighted frequently used terms in each category, allowing better interpretation of patterns in abusive language specific to Tamil and Malayalam.

3.4 Preprocessing

The text becomes more uniform after being converted into lowercase, having all the punctuation signs stripped off, and numbers excluded. Words were rearranged to cut them down to their root forms or lemmatized but with meaning in various forms of the word.

To convert text to numerical features, we employed vectorization using Term Frequency-Inverse Document Frequency (TF-IDF), which is a method that analyzes the relevance of a certain word within a certain document in relation to the entire data collection available. Term Frequency (TF) is the frequency of how often a word is found in a document, and Inverse Document Frequency (IDF) scales down the value of frequently occurring words so that highly used but less informative words do not dominate the model.

We used TF-IDF with unigrams and bigrams, with unigrams as single words and bigrams as two-word sequences, preserving the contextual relationship between words. This representation is more effective for the model to detect patterns of abusive language.

3.5 Machine Learning Models

To classify Tamil and Malayalam comments as abusive or non-abusive, we examined and evaluated three different machine learning models in order to compare them. Each model has been selected on the basis of suitability for efficient text data processing and ability to handle the nuances of classification tasks.

- **Logistic Regression:** A probabilistic model that predicts the probability that comments are abusive using the logistic function. Vectored features representing the TF-IDF were used as input, and hyperparameters such as regularization strength C and solver optimization were tuned through grid searches. It provides a strong baseline performance with an accuracy of 87.48%, offering simple and interpretable results.
- **Support Vector Machine (SVM):** SVM uses the optimal hyperplane to separate abusive and non-abusive comments in the high-dimensional TF-IDF space. With a linear kernel, it identifies the optimal hyperplane. SVM achieved the highest accuracy (89.89%), demonstrating robust handling of text data and effective generalization.

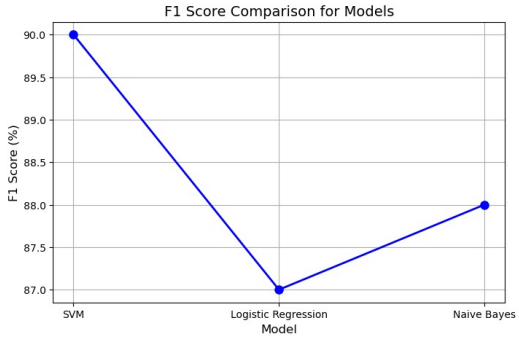


Figure 2: F1 score comparison for 3 models: Support Vector Machine, Logistic Regression, Naive Bayes.

- **Naive Bayes:** This probabilistic classifier uses Multinomial Naive Bayes to evaluate the likelihood of words contributing to each class. Its simplicity and effectiveness make it ideal for quick training and testing, achieving an accuracy of 87.51%.

3.6 Model Evaluation

The models were evaluated using metrics such as accuracy, F1 score, precision, and recall to ensure a balanced classification of offensive and non-abusive comments. Logistic regression achieved an accuracy of 87.48% and an F1 score of 87%, providing good baseline performance. SVM outperformed other models with the highest accuracy of 89.89% and an F1 score of 90%, showing its robustness in handling high-dimensional text data. As a result, Naive Bayes achieved an accuracy of 87.51% and an F1 score of 87%, known for its efficiency. Cross-validation was used to ensure robustness; the F1 score was prioritized to balance precision and recall. These evaluations demonstrate that while SVM achieved the best overall performance, logistic regression and Naive Bayes provide reliable and efficient

alternatives for deployment. As illustrated in Figure 3., enhanced model performance Improvement will be guided by an analysis of the 174 false positives, maybe using methods such as weighted loss functions for class imbalance. Plots like precision-recall curves and ROC will also be useful. Our system will be better able to trust people and make wise decisions when identifying abusive content if we reduce false positives.

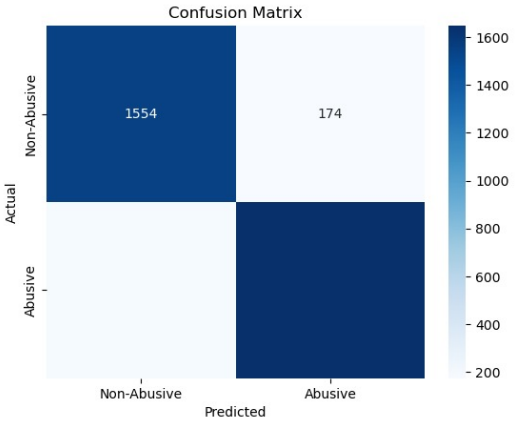


Figure 3: Comprehensive Evaluation and Improvement for the Malayalam dataset. Visualizing the model's performance in terms of false positives and false negatives.

4 Results

The project results indicate the effectiveness of different machine learning models for classifying abusive and non-abusive comments in Tamil and Malayalam. The models were used for evaluation based on accuracy, F1 score, precision, and recall so that the models exhibit almost balanced performance in both class categories (abusive and non-abusive). Here, a detailed summary of results is presented in Table 2.

The results provide strong evidence for the detection of abusive content in Tamil and Malayalam. It is advised that SVM be used for deployment due to its efficient performance on all fronts. Logistic regression and Naïve Bayes can act as simpler workarounds subject to resources. This system promises much greater potential use on real-time social media content management and user protection.

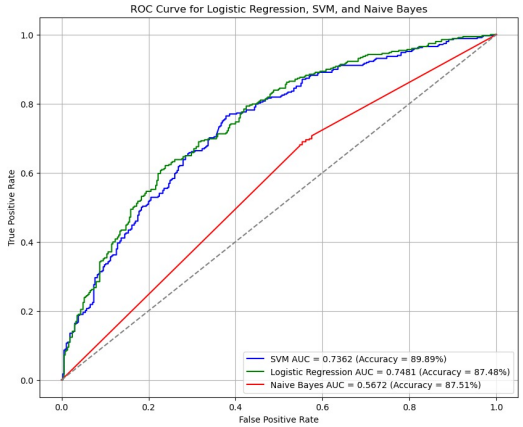


Figure 4: The ROC curve shows that SVM performs the best, followed by Logistic Regression, with Naive Bayes having the lowest AUC, indicating relatively poorer performance.

5 Conclusion

The widespread existence of gender-based abuse on social media platforms requires effective mechanisms to detect and mitigate abusive content targeted at women. In this work we have attempted the challenge of classifying comments in Tamil and Malayalam as abusive or non-abusive by applying robust preprocessing, exploratory analysis, and machine learning techniques. By using TF-IDF vectorization and fine-tuning hyperparameters, three models are Logistic Regression, Support Vector Machine (SVM), and Naive Bayes.

Among them, SVM was found to be the best model with an overall accuracy of 89. 89% and balanced F1 scores of 0. 90% for both the Abusive and Non-Abusive classes. This goes to show how SVM handles high-dimensional text features effectively, ensuring fair detection across categories. Among the other statistical procedures evaluated were Naive Bayes and Logistic Regression. The proposed solution combines high accuracy with interpretability by using WordCloud visualization to gain insight into language patterns. The study shows that machine learning can be used for automated moderation of abusive comments in Tamil and Malayalam languages. The results show the feasibility of using machine learning for automated moderation of abusive comments in Tamil and Malayalam-speaking language environments.

6 Limitation

The proposed solution was able to effectively detect abusive comments in both Tamil and Malayalam; still, a couple of limitations arose that would eventually further improve the model. It is not endowed with the deep contextual understanding that traditional models usually rely on, the basis of SVM, logistic regression, and naïve Bayes, as those models tend to use TF-IDF features and fail to incorporate implicit abuse, sarcasm, and

Model	Accuracy (%)	F1-Score (%)	Precision (Class 0, 1)	Recall (Class 0, 1)	AUC-ROC
SVM	89.89	90	(0.89, 0.90)	(0.90, 0.90)	0.7362
Logistic Regression	87.48	87	(0.88, 0.87)	(0.86, 0.89)	0.7481
Naive Bayes	87.51	88	(0.90, 0.86)	(0.84, 0.91)	0.5672

Table 2: Comparison of Logistic Regression, SVM, and Naive Bayes models on accuracy, F1-score, precision, recall and AUC-ROC curve for classifying abusive and non-abusive comments.

code-mixed language. This is also true because it only serves to process texts, and, most of the time, social media abuse on the actual platforms would encompass multimodal media like images, memes, or videos. It simply cannot sense the visual characteristics of the input, which, in turn, leads to non-recognition of abusive content in image formats or any other multimedia types. Another such limitation is in the context-based elements of any conversation, as this treats comments to be treated and looked into entirely independently without seeking prior interaction among them, leading it to be unable to identify indirect and unfolding abuse within several messages.

The training set is linguistically and culturally so limited in breadth that the current model does not allow it to better generalize to dialectic and other variant forms of spoken Tamil and Malayalam. Another challenge with biased training data is the problem of unfair classification, which hurts specific user groups. Finally, deep learning models for real-time deployment pose challenges in terms of computation: very fast inference with little loss in accuracy remains an open question.

6.1 Future work

To overcome these limitations, the next wave of future work focuses on some areas of improvement. The upgraded transformer-based models with BERT, mBERT, and IndicBERT will be widely used to improve contextual understanding and classification accuracy. The multimodal abuse detection module will be integrated by looking into the textual and visual aspects, where the system can find the harmfulness of content in a meme or other multimedia formats. This way, it can see more subtle, context-dependent abuse once it has gained the capacity for processing many threads and interactions into its historical analysis. This increases the size of the dataset over variations in combinations of linguistic or cultural differences that further augment the methods of data augmentation. Back-translations and replacement of synonyms, among others, may generalize a model for many Dravidian languages. Model optimization techniques like quantization, pruning, and knowledge distillation will be used to reduce the computational overhead while keeping the accuracy intact. Strategies for bias mitigation using explainable AI will be applied to the system to make it fairer and more interpretable in terms of responsible and ethical AI. The final aspect is cross-lingual transfer learning, which would enable the system to support multiple Dravidian languages, thereby making it applicable to a large extent. User feedback with mechanisms involving adaptive learning, where the system keeps on improving continuously. Adaptability toward the newly emerging patterns of online abuse would enable it to track these events correctly. So, with such updates and improvements, this proposed system could be a little more robust in terms of efficiency, as well as just fair enough regarding the detection of the right abusive content.

Acknowledgment

We thank DravidianLangTech-2025 at NAACL 2025 shared task organizers for providing data sets and guidance.https://sites.google.com/view/dravidianlangtech-2025/shared-tasks-2025

References

Fazlourrahman Balouchzahi, Anusha Gowda, Hosa Halli Shashirekha, and Grigori Sidorov. 2022. MUCIC@TamilNLP-ACL2022: Detection of abusive comments in Tamil using 1D Conv-LSTM. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213.

Bharathi Raja Chakravarthi, et al. 2021. Findings of the Shared Task on Offensive Language Identification in Tamil, Malayalam and Kannada. In *DRAVIDIANLANGTECH*, pages 1–10.

Rajalakshmi, Ratnavel, Duraphe, Ankita, Shibani, Antonette. 2022. Abusive comment detection in Tamil using multilingual transformer models. *DLRG@DravidianLangTech-ACL2022*, 207–213.

Kathiravan Pannerselvam, et al. 2023. CSSCUTN@DravidianLangTech: Abusive comments detection in Tamil and Telugu. *DRAVIDIANLANGTECH*, pages 1–15.

Zichao Li. 2021. Codewithzichao@DravidianLangTech-EACL2021: Exploring multilingual transformers for offensive language identification on code-mixing text. *DRAVIDIANLANGTECH*, pages 100–110.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Siva Sai and Yashvardhan Sharma. 2021. Towards offensive language identification for Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 18–27, Kyiv.

R Anita and CN Subalalitha. 2019a. An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4. IEEE.

Judith Jeyafreeda Andrew. 2021. JudithJeyafreedaAndrew@DravidianLangTechEACL2021: Offensive language detection for Dravidian code-mixed YouTube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 169–174, Kyiv.

CN Subalalitha and E Poovammal. 2018. Automatic bilingual dictionary construction for Tirukural. *Applied Artificial Intelligence*, 32(6):558–567.

Pradeep Kumar Roy and Abhinav Kumar. 2021. Sentiment analysis on Tamil code-mixed text using BiLSTM. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation*, pages 100–110, Online. CEUR.

Fazlourrahman Balouchzahi, Anusha Gowda, Hosa Halli Shashirekha, and Grigori Sidorov. 2022. MUCIC@TamilNLP-ACL2022: Detection of abusive comments in Tamil using 1D Conv-LSTM. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213.

B Bharathi and A Agnusimmaculate Silvia. 2021. SSNCSE NLP@DravidianLangTech-EACL2021: Offensive language identification on multilingual code-mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 313–318, Kyiv.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the Shared Task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing, September.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of Abusive Comment Detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics. https://aclanthology.org/2022.dravidianlangtech-1.44, DOI: 10.18653/v1/2022.dravidianlangtech-1.44.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Association for Computational Linguistics.

Celestia@DravidianLangTech 2025: Malayalam-BERT and m-BERT based transformer models for Fake News Detection in Dravidian Languages

Syeda Alisha Noor[1], Sadia Anjum[2], Syed Ahmad Reza[1], and Md. Rashadur Rahman[1]

[1]Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
[2]International Islamic University Chittagong, Chattogram, Bangladesh
u1904018@student.cuet.ac.bd, sadiaanjum210@gmail.com, u1904016@student.cuet.ac.bd
rashadur@cuet.ac.bd

Abstract

Fake news detection in Malayalam is difficult due to limited data and language challenges. This study compares machine learning, deep learning, and transformer models for classification. The dataset is balanced and divided into training, development and test sets. Machine learning models (SVM, Random Forest, Naive Bayes) used TF-IDF features and deep learning models (LSTM, BiLSTM, CNN) worked with tokenized sequences. We fine-tuned transformer models like IndicBERT, MuRIL, mBERT, and Malayalam-Bert. Among them, the Malayalam-Bert model performed the best and achieved an F1 score of 86%. On the other hand mBERT performed best at spotting fake news. However, the models struggled with mixed-language text and complex writing. Despite these challenges, transformer models turned out to be the most effective for detecting fake news in Malayalam.

1 Introduction

Fake news is spreading fast on the internet, and it has become important to find better ways to detect it, especially in languages that do not have many digital resources. The shared task on fake news detection in Dravidian languages, held at DravidianLangTech@NAACL 2025, was organized to tackle this problem. This research focused on developing and testing different methods to identify fake news in Dravidian languages. Since these languages are complex, the task used special approaches instead of general models.

Fake news detection (FND) can be divided into two types: monolingual and multilingual. Monolingual FND is used to find fake news in one language. On the other hand, multilingual FND is needed when fake news is mixed with multiple languages, including code-mixed content. Detecting fake news in low-resource languages is difficult because there is a lack of enough labeled datasets, pre-trained models, or other digital tools. However, some methods like collecting and labeling data, using cross-lingual models, applying transfer learning, and creating models suited for specific languages can help to improve fake news detection.

In this study, we tested four pre-trained transformer models, Indic-BERT, m-BERT, Malayalam and MuRIL, to determine whether transfer learning from high-resource languages could improve fake news detection in Dravidian languages. Their findings demonstrated the feasibility of this approach, highlighting the role of advanced NLP techniques in mitigating misinformation in underrepresented languages. The implementation details have been provided in the following GitHub repository:- https://github.com/Alisha1904018/Share-task-2025.

2 Related Work

Fake news detection has become a growing area of research nowadays. Low-resource languages like Dravidian languages form a significant field of study. Many studies have focused on using machine learning and deep learning approaches to address this problem.

(Raja et al., 2024) developed a hybrid model that combines CNN and BiLSTM to detect fake news in Dravidian languages. They used MuRIL to obtain better language-specific details and reduce overfitting. Their model performed better than state-of-the-art approaches.

(Shanmugavadivel et al., 2024) also took part in DravidianLangTech 2024 and experimented with machine learning models such as Random Forest, Logistic Regression, and Decision Trees.

(Farsi et al., 2024) customized a MuRIL-BERT model and evaluated it with different machine learning and transformer-based techniques. Their strategy resulted in an F1-score of 0.86 for binary classification and 0.5191 for multi-class classification.

(Shohan et al., 2024)achieved the highest F1 scores of 75.82% by using RoBERTa for English tweets in classifying check worthy sentences.

(Rahman et al., 2024) achieved the highest macro F1-score (0.88) in Malayalam fake news detection using Malayalam-BERT, securing the top position in the shared task.

(Osama et al., 2024) explored both machine learning models, such as SVM, Random Forest, Logistic Regression, and Naïve Bayes, and deep learning models: CNN, BiLSTM, and BiLSTM with attention, along with transformers. The best-performing model in this study was m-BERT, which had an F1-score of 0.85 and ranked 4th in the shared task.

(Borgohain et al., 2023) created a dataset named Dravidian_Fake, containing 26,000 fake news articles in Telugu, Tamil, Kannada, and Malayalam languages. This study has attained the highest accuracy of 93.31% with mBERT and XLM-R using adaptive fine-tuning for multilingual fake news classification.

(Raja et al., 2023) tested four transformer models—mBERT, XLM-RoBERTa, IndicBERT, and MuRIL—on Telugu, Kannada, Tamil, and Malayalam fake news detection. Among these, MuRIL performed the best.

(Yigezu et al., 2024) introduced *Ethio-Fake*, a framework that integrates social-contextual and content-driven attributes for misinformation detection in low-resource languages. They evaluated various techniques, including traditional machine learning, neural networks, and transfer learning, concluding that ensemble learning achieved the highest F1-score of 0.99.

(Wang et al., 2024) conducted an extensive survey on monolingual and multilingual misinformation detection for under-resourced languages. Their work reviewed existing datasets, methodologies, and challenges in the field, emphasizing the need for improved data collection and inclusive AI strategies. They also highlighted the effectiveness of language-agnostic and multi-modal approaches in combating misinformation.

(Shimi et al., 2024) focused on *language identification* for Dravidian languages, a crucial step in fake news detection within multilingual settings. They compared machine learning and deep learning models for recognizing languages like Tamil and Malayalam. Their results indicated that deep learning-based language-independent models achieved the highest accuracy of 98%.

3 Task and Dataset Description

Fake News Detection in Dravidian Languages comprises balanced and normalized data (Subramanian et al., 2025) given by the organizers (Subramanian et al., 2024), basically aiming at building some systems (Devika et al., 2024) to label a original versus fake news of the Malayalam language (Subramanian et al., 2023) posts found in the media.The dataset consists of 3 portion: train, test & dev dataset

Class	Train	Development	Test
Fake	1599	406	507
Original	1658	409	512
Total	3257	815	519

Table 1: Dataset analysis

4 Methodology

The proposed method is experimented by using different machine learning, deep learning and transformer-based approach to classify fake news in a code-mixed Malayalam-English dataset. Our approach consists of data preprocessing, feature extraction, model development, and performance evaluation.

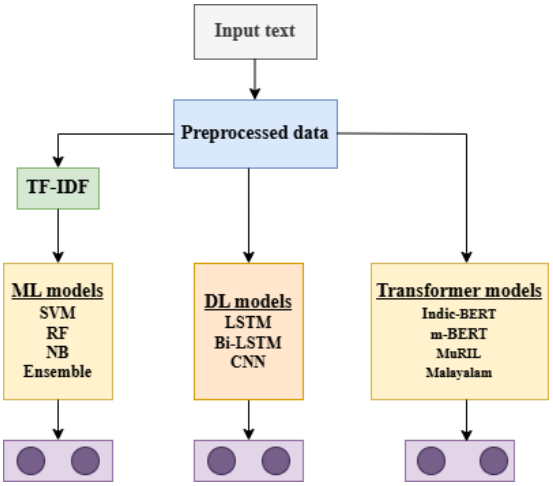


Figure 1: Methodological outline

4.1 Data preprocessing

As the dataset contains code-mixed text, proper preprocessing was crucial for ensuring meaningful feature representation. We have conducted several steps to achieve this. At first, we performed text cleaning was performed by removing emoticons, pictographs, URLs, and stopwords to eliminate noise. Next, to handle code-mixed text,

we used the Indic-Transliteration library to convert English words to Malayalam, ensuring consistency in linguistic representation. Finally, we employed subword-based tokenization for deep learning and transformer models to effectively process the mixed-language text.

4.2 Machine Learning Approaches

For the classification of the model, we used logistic regression (LR), support vector machine (SVM), random forest (RF), decision tree (DT) and naive bayes (NB). We used scikit-learn[1] to tune each model: max iterations of 1000 for LR, SVM with radial basis function(rbf) kernel, 1000 estimators for RF, unlimited depth for DT, and an alpha of 0.15 for NB. Majority voting with the ensemble model combined the LR, SVM, and DT models to increase robustness. The features were extracted using TF-IDF.

4.3 Deep Learning Approaches

We perform fake news classification using the LSTM, BiLSTM, and CNN models. The model development was done on the 'TensorFlow' framework. The text data has been tokenized and then converted into a padded sequence with a vocabulary size of 10,000 and a sequence length of 100. The LSTM model consists of two LSTM layers with 128 and 64 units, respectively. Each is followed by a 0.3 dropout layer to handle overfitting. The BiLSTM model consisted of two bidirectional LSTM layers of 128 and 64 units, respectively, using the same dropout strategy. The CNN model consisted of two 1D convolutional layers with 128 and 64 filters, kernel size 5, and max-pooling layers to capture the feature set. All of them used the same embedding layer of dimension 128, the Adam optimizer, and binary cross-entropy loss. The results were best after training up to a certain 10 epochs with a batch size of 32. Model performance was evaluated based on accuracy and F1 score.

4.4 Transformer Approaches

We fine-tuned pre-trained multilingual transformer models such as IndicBERT(Deode et al., 2023), MuRIL (Khanuja et al., 2021), mBERT (Devlin et al., 2019), and Malayalam (Joshi, 2023) from the Hugging Face[2] transformers library for our fake news classification task. For compatibility with each model, AutoTokenizer was instantiated with a sequence length of 256 tokens. We have used a batch size of 16, a learning rate of $2e^{-5}$, and trained for 10 epochs. We employed the Adam optimizer for gradient descent and used cross-entropy loss. We've trained the model using the dataset, and we've been evaluating its performance using accuracy and the classification report to measure its effectiveness in classifying fake news. Among all the model Malayalam has shown superior performance than all the other models.

Hyperparameter	Value
Batch Size	16
Optimizer	Adam
Epochs	10
Dropout Rate	0.3
Learning Rate	$2e^{-5}$

Table 2: Hyperparameter tuning

5 Result Analysis

This section presents a comparative performance analysis of various experimental approaches. The efficiency of the models is primarily assessed based on the F1-score, while precision and recall are also considered in some cases. A summary of the precision (P), recall (R), and F1-score for each model on the test set is presented in Table 3. Table 3 illustrates that for fake news classification, Malayalam perform the best. Among ML model, SVM, NB & Ensemble(SVM + DT + Logistic Regression) shows the same result. Among DL models, BiLSTM & CNN shows almost same result outperforming LSTM. Among Transformer, MuRIL shows a superior performance than any other models.

Method	Classifier	P	R	F1
ML	SVM	0.75	0.75	0.75
	RF	0.72	0.72	0.72
	NB	0.75	0.75	0.75
	Ensemble	0.75	0.75	0.75
DL	LSTM	0.25	0.50	0.33
	BiLSTM	0.81	0.81	0.81
	CNN	0.82	0.82	0.82
Transformers	Indic-BERT	0.76	0.75	0.74
	m-BERT	0.84	0.84	0.84
	MuRIL	0.83	0.82	0.82
	Malayalam	0.86	0.86	**0.86**

Table 3: Comparative analysis of performance on test data. Here P, R & F represent precision, recall & F1 score, respectively.

[1] https://scikit-learn.org
[2] https://huggingface.co/

6 Error Analysis

The misclassification occurred due to multiple challenges in the model's interpretation. It struggled with indirect speech and quotes. It misidentified sentences with numerical references, assuming numbers indicate factual accuracy. It also failed

Text	Actual label	Predicted label
ഓഷോ രജനീഷ് പറഞ്ഞപോലെ എനിക്കപ്പോൾ തോന്നിയത് അങ്ങനെയാണ് ഇപ്പോൾ തോന്നുന്നത് ഇങ്ങനെയാണ് എന്തൊക്കെയോ ആവോ	Fake	Original
ഫസ്റ്റ് ചോദ്യേ ജനഘല് അന്ധു വിദ്ദേ ബുസിനെസ്സ് ചേയ്ഥലുമ് വിജയികുമ് ഇന് ഇന്ദിഅ	Fake	Original
വ്ഹ്യ് അല്ല് ചോഉന്തിഏസ് ച്ലുബ് തോഗേഥേർ ന് ബ്ലോച്ച് ചിന തോ ഥേ മക്ഷിമുമ് ഏക്ഷ്തേന്ദ് അന്ദ് രേനമേ ചോവിദ് ഥൻ ചിനേസേ വിരുസ് ഇന് അല്ല് സുച് ചസേസ് വ്ഹേന് സോമേ ഓനേ വോർക് തോ ദേസ്തോയ് ഓഥേർസ് ഥേ നതുരേ ഇന്തേരേനേ ന് ചോന്തൈന് സുച് നേഗതിവേ ഭോർചേ ഥത് ഇസ് ഥേ ഗ്രചേ ഓഫ് ഗോദ് ഥേ ത്രുഥ്	Fake	Original

Figure 2: Misclasssification of text

to detect complex structures and misleading tones, mistaking deceptive content for original. These issues highlight the model's difficulty in recognizing nuanced patterns of misinformation. From Figure 3 and Figure 4, it can be seen that the Malayalam-Bert model correctly identified 383 fake news samples as fake and 468 original news samples as original. On the other hand, the mBERT model correctly

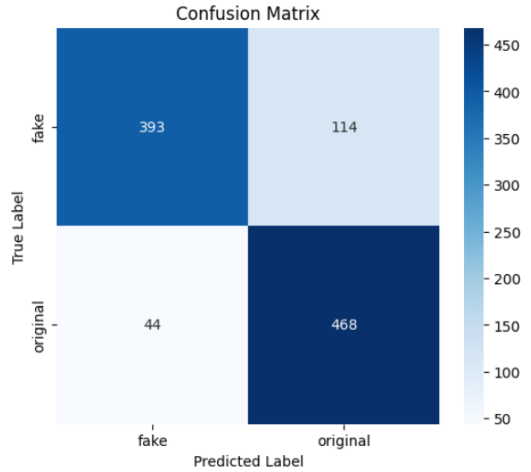


Figure 3: Confusion matrix of the Malayalam-BERT model

identified 427 fake news samples as fake and 428 original news samples as original. The confusion matrix of the top-performing

models (Malayalam and mBERT) is displayed in Figure 2, highlighting the highest precision achieved by the Malayalam model, as it correctly classifies most of the samples. Malayalam correctly identified 393 fake samples out of 507, whereas mBERT correctly identified 427 fake samples. Since mBERT successfully classifies more fake

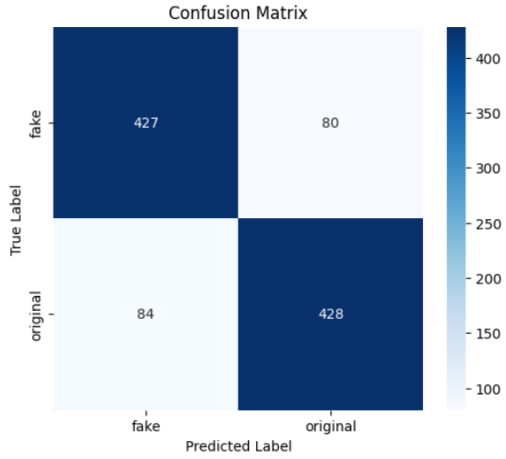


Figure 4: Confusion matrix of the m-BERT model

news samples, it can be concluded that mBERT performs better in fake news detection. However, the Malayalam model achieved the highest accuracy.

7 Conclusion

This paper discusses the detection of fake news in Dravidian languages by evaluating various ML, DL, and transformer-based approaches. Our experimental results document the best performance to be that of the Malayalam-BERT model with a maximum F1-score of 0.86 among the considered approaches. It further strengthens the efficiencies of transformer-based architectures which can handle complex linguistic structures in low-resource languages. In addition, future studies can be conducted by improving the performance of data augmentation, hybrid modeling techniques, and ensembling multiple transformer-based models to further improve robustness in fake news detection.

Limitations

While our approach demonstrates better performance, it has certain limitations also

- As Malayalam is a low-resourced language, it is difficult to capture its inherent linguistic complexities.
- Due to resource constraints, transformer ensembling could not be performed.

References

Samir Borgohain, Badal Soni, and Eduri Raja. 2023. Fake news detection in dravidian languages using

transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Salman Farsi, Asrarul Eusha, Ariful Islam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech EACL2024: Fake news detection in Malayalam language leveraging fine-tuned MuRIL BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, St. Julian's, Malta. Association for Computational Linguistics.

Raviraj Joshi. 2023. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *Preprint*, arXiv:2211.11418.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_GoodFellows@DravidianLangTech EACL2024: A transformer-based approach for detecting fake news in Dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, St. Julian's, Malta. Association for Computational Linguistics.

Tanzim Rahman, Abu Raihan, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@DravidianLangTech EACL2024: Fake news classification using Malayalam-BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, St. Julian's, Malta. Association for Computational Linguistics.

Eduri Raja, Badal Soni, and Sami Kumar Borgohain. 2024. Fake news detection in dravidian languages using multiscale residual cnn_bilstm hybrid model. *Expert Syst. Appl.*, 250:123967.

Eduri Raja, Badal Soni, and Samir Borgohain. 2023. *Fake News Detection in Dravidian Languages Using Transformer Models*, pages 515–523.

Kogilavani Shanmugavadivel, Malliga Subramanian, Sanjai R, Mohammed Sameer B, and Motheeswaran K. 2024. Beyond tech@DravidianLangTech2024 : Fake news detection in Dravidian languages using machine learning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, St. Julian's, Malta. Association for Computational Linguistics.

G. Shimi et al. 2024. Language identification for dravidian languages: A crucial step for fake news detection in multilingual settings. *TBD*.

Symom Shohan, Md Hossain, Ashraful Paran, Shawly Ahsan, Jawad Hossain, and Moshiul Hoque. 2024. Semanticcuetsync at checkthat! 2024: Pre-trained transformer-based approach to detect check-worthy tweets.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Xinyu Wang et al. 2024. A survey on monolingual and multilingual misinformation detection for low-resource languages. *TBD*.

Mesay Gemeda Yigezu et al. 2024. Ethio-fake: Integrating social-contextual and content-based features for fake news detection in under-resourced languages. *TBD*.

Trio Innovators @ DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages

Radha N, Swathika R, Farha Afreen I, Annu G, Apoorva A
Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, India
radhan@ssn.edu.in
swathikar@ssn.edu.in
farha2110729@ssn.edu.in
annu2110538@ssn.edu.in,apoorva2110445@ssn.edu.in

Abstract

This paper presents an in-depth study on multimodal hate speech detection in Dravidian languages—Tamil, Telugu, and Malayalam—by leveraging both audio and text modalities. Detecting hate speech in these languages is particularly challenging due to factors such as code-mixing, limited linguistic resources, and diverse cultural contexts. Our approach integrates advanced techniques for audio feature extraction and XLM-Roberta for text representation, with feature alignment and fusion to develop a robust multimodal framework. The dataset is carefully categorized into labeled classes: gender-based, political, religious, and personal defamation hate speech, along with a non-hate category. Experimental results indicate that our model achieves a macro F1-score of 0.76 and an accuracy of approximately 85

1 Introduction

Hate speech on social media is a significant issue, particularly in underrepresented Dravidian languages like Tamil, Telugu, and Malayalam, where linguistic diversity and limited resources pose challenges to detection. Traditional text-based models often fail to capture crucial audio cues like tone and emotion, limiting their effectiveness.

This study introduces a multimodal approach that integrates both text and audio to enhance detection accuracy. A real-world dataset is used, incorporating labeled text and audio across multiple hate speech categories, including gender, politics, religion, and personal defamation, as well as non-hate speech. Audio data helps identify subtle cues like sarcasm, aggression, and emphasis, which are often lost in text-only models.

The methodology employs Wav2Vec 2.0 for extracting speech features and XLM-Roberta for multilingual text embeddings. These features are aligned, fused, and classified using XGBoost, achieving a macro F1-score of 0.76, demonstrating the model's robustness in handling multimodal and multilingual data.

This research highlights the value of integrating multiple modalities for hate speech detection in low-resource languages. Future work will focus on expanding the dataset, incorporating additional modalities like video for better context understanding, and refining models with contextual embeddings and domain-specific fine-tuning. This lays the foundation for developing more effective hate speech detection systems, fostering a safer and more inclusive online environment.

2 Literature Review

Hate speech detection has been extensively studied across languages, platforms, and contexts using various machine learning and deep learning techniques. Researchers have explored NLP methods, multimodal approaches, and resource-specific challenges to enhance detection accuracy. Key challenges include dataset inconsistencies, linguistic nuances, and multimodal data integration.

Several studies analyze different approaches to hate speech detection. (Alkomah and Ma, 2022) emphasize the need for larger, more diverse datasets and improved feature selection due to dataset inconsistencies. (Fortuna and Nunes, 2018) highlight the limitations of basic word filters, advocating for sophisticated NLP techniques that consider linguistic context and multimodal data. (Jahan and Oussalah, 2023) review NLP techniques, discussing machine learning models, feature extraction, and challenges like contextual understanding.

Language-specific research has advanced hate speech detection in low-resource settings. (Parker and Ruths, 2023) introduce the OptimizePrime model for Tamil, surpassing existing methods.(Roy et al., 2022) propose a deep ensemble framework for Tamil, Malayalam, and Kannada, emphasizing

linguistic features and context. (Bansod, 2023) explores Hindi hate speech detection, highlighting linguistic and cultural influences, while (Sutejo and Lestari, 2018) improve Indonesian hate speech detection using deep learning. (Li, 2021)suggests methods to address challenges in low-resource settings, such as small datasets and limited computational resources.

The rise of hate speech on social media has increased interest in multimodal approaches. (Gomez et al., 2019) demonstrate that integrating text, images, and videos enhances detection accuracy. (Wu and Bhandary, 2020) apply machine learning to detect hate speech in videos using speech recognition and visual context analysis.(Toliyat et al., 2022) analyze the surge in pandemic-related racial hostility against Asians, evaluating NLP-based detection methods.(Haque and Chowdhury, 2023) use ensemble learning to improve detection robustness.

Comparative studies refine deep learning-based hate speech detection. (Abro et al., 2020) compare machine learning algorithms, analyzing their strengths and weaknesses.(Malik et al., 2022) examine deep learning techniques to identify optimal models for classification.(Zhou et al., 2021) enhance precision by integrating sentiment analysis into detection models. (Haque and Chowdhury, 2023) demonstrate that ensemble learning improves performance and reliability.

Beyond technical advancements, ethical concerns in hate speech detection have been examined. (Parker and Ruths, 2023) assess biases, limitations, and societal impacts of automated systems. (Kovács et al., 2021) explore strategies to address data scarcity in social media hate speech detection through external data sources and improved model generalization. Collectively, these studies contribute to evolving methodologies, ensuring accuracy, efficiency, and ethical considerations in hate speech detection.

3 Proposed Methodology

3.1 Dataset Description

The dataset represents real-world hate speech scenarios in Tamil, Telugu, and Malayalam, comprising 300 samples equally divided among the three languages, with 50 text and audio samples each for training and testing. It is structured for multimodal hate speech classification across five categories: Gender-based Hate (G), Political Hate (P), Religious Hate (R), Personal Defamation (C), and Non-Hate Speech (NH). The training dataset includes 407 Tamil, 440 Telugu, and 706 Malayalam samples, while the validation dataset consists of 102 Tamil, 111 Telugu, and 177 Malayalam samples. The test dataset remains balanced with 50 samples per language. Each file follows a standardized naming format, such as H_MŁ_001_C_F_044_001.WAV, where "H" indicates hate speech, "ML" represents the language (Malayalam), "F" refers to the speaker's gender (Female), "044" is the source video identifier, and "001" is the utterance number. The dataset presents challenges such as class imbalance, where Non-Hate Speech dominates while Personal Defamation is underrepresented. Code-mixing is another complexity, with text containing both native and English scripts, reflecting real-world social media usage. Additionally, variations in tone, pitch, and speaking styles in audio data impact feature extraction. By integrating multimodal data across different languages and real-world scenarios, this dataset provides a comprehensive foundation for developing and evaluating hate speech detection models.

3.2 Audio and Text Feature Extraction

TThe Wav2Vec 2.0 model is used for audio feature extraction, with audio resampled to 16 kHz using Librosa for compatibility. The feature vector, A, derived from the model's final hidden state, is formulated as: $A = Wav2Vec2(x)$, where x is the resampled input. Extracted features capture key acoustic properties: tone (indicating aggression or sarcasm), pitch (high variations signaling excitement or anger), intensity (reflecting emphasis), and **speech rhythm/duration** (detecting elongated or stressed syllables). Additional features include MFCCs, spectral contrast (capturing energy variations), zero-crossing rate (ZCR) (indicating abrupt sound changes), and log Mel spectrogram (representing energy distribution across frequencies), enhancing the detection of hateful speech.

Dataset	Tamil	Telugu	Malayalam
Training	407	440	706
Validation	102	111	177
Test	50	50	50

Table 1: Dataset Distribution for Tamil, Telugu, and Malayalam

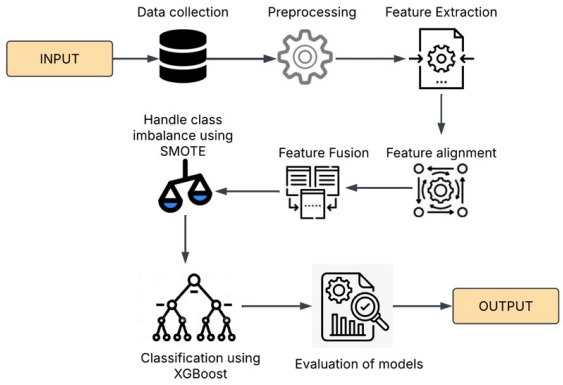


Figure 1: Proposed Architecture

Text data is processed using XLM-RoBERTa, where the [CLS] token embedding, T, is extracted from the transformer's final layer:

$$T = \text{XLM-RoBERTa}(text) \tag{1}$$

This embedding captures the semantic meaning of the text, which is crucial for understanding context and intent. Preprocessing involves cleaning the text by removing punctuation and unwanted characters using regular expressions. The extracted embeddings incorporate contextual meaning, identifying sentiment or hatefulness, and code-mixing patterns, leveraging XLM-RoBERTa's multilingual capabilities for handling Tamil-English, Telugu-English, or Malayalam-English text. Additionally, the model emphasizes keywords commonly linked to hate speech, extracts word embeddings to capture relationships and context, and derives sentence-level embeddings from the [CLS] token to represent overall text semantics effectively.

3.3 FeatureAlignment

Features from both audio and text modalities are aligned based on file names to ensure consistency within the dataset. Each instance in the dataset, represented as D_i, consists of the corresponding audio and text feature vectors:

$$D_i = (A_i, T_i) \tag{2}$$

where A_i represents the extracted audio features, and T_i denotes the text embeddings. This alignment ensures that each data instance contains synchronized multimodal information for effective hate speech detection.

3.4 Feature Fusion

Once the features are aligned, they are fused by horizontally concatenating the audio and text embeddings to form a unified feature representationF_i

$$F_i = [A_i \parallel T_i] \ \setminus \tag{3}$$

where || denotes the concatenation operation. This fused representation allows the model to leverage both acoustic signals (such as tone, pitch, and intensity) and semantic information (such as contextual meaning, sentiment, and keyword emphasis) for classification, enhancing its ability to detect hateful speech more effectively.

3.5 Handling Class Imbalance

To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied. SMOTE generates synthetic samples in the feature space for the minority class, ensuring a balanced dataset. The synthetic samples S are created using the function

$$S = G(C_{\text{minority}}) \tag{4}$$

where G represents the SMOTE algorithm and C_minority denotes the original minority class samples. This approach prevents bias in classification by ensuring that the model does not favor the majority class.

3.6 Classification Model

For classification, the XGBoost model, known for its robustness in handling imbalanced datasets, is employed. The model takes the fused feature vector F as input and predicts the hate speech category y as follows

$$y = \text{XGBoost}(F_i) \tag{5}$$

By leveraging gradient boosting, feature importance weighting, and optimized decision trees, XGBoost effectively learns patterns from both audio and text modalities, ensuring accurate hate speech detection.

4 Experiment and Results

An XGBoost classifier was trained using fused feature vectors, with the training process optimized to address class imbalance through the application of SMOTE, ensuring fair representation of minority classes. To enhance performance, hyperparameter tuning was conducted, setting the learning rate

to 0.01, the maximum depth to 6, and the number of estimators to 1000. For each language, the model's performance was evaluated using text features, combined features, and audio features. Tables below presents the classification report and validation accuracy results for each feature set.

Category	Precision	Recall	F1-Score	Support
C	0.00	0.00	0.00	13
G	0.20	0.08	0.11	13
N	0.55	0.72	0.63	58
P	0.00	0.00	0.00	7
R	0.12	0.08	0.10	12

Table 2: Accuracy of Tamil audio dataset

The macro-average F1-score is 0.17, while the weighted average F1-score is 0.38 (refer Table 2)

Category	Precision	Recall	F1-Score	Support
C	0.15	0.17	0.16	12
G	0.66	0.83	0.73	58
N	0.33	0.08	0.13	12
P	0.33	0.14	0.20	7
R	0.60	0.46	0.52	13

Table 3: Accuracy of Tamil audio-text dataset

The macro-average F1-score is 0.35, while the weighted average F1-score is 0.53.(refer Table 3)

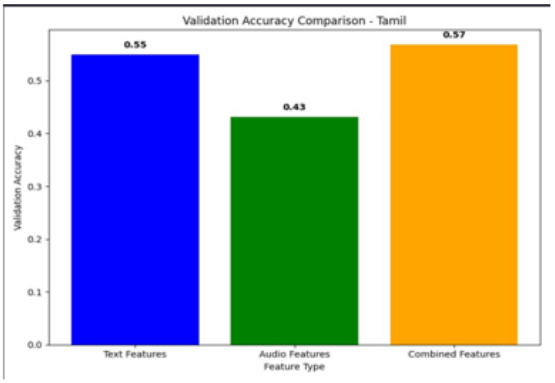


Figure 2: Validation Accuracy Comparison for Tamil Using Different Feature Types

The macro-average for malayalam text F1-score is 0.43, while the weighted average F1-score is 0.55.

Category	Precision	Recall	F1-Score	Support
C	0.16	0.16	0.16	37
G	0.12	0.12	0.12	17
N	0.48	0.52	0.50	81
P	0.09	0.08	0.09	24
R	0.00	0.00	0.00	18

Table 4: Accuracy of Malayalam audio dataset

The macro-average F1-score is 0.17, while the weighted average F1-score is 0.29 (refer Table 4)

Category	Precision	Recall	F1-Score	Support
C	0.68	0.68	0.68	37
G	0.63	0.75	0.69	81
N	0.29	0.22	0.25	18
P	0.27	0.17	0.21	24
R	0.21	0.18	0.19	17

Table 5: Accuracy of Malayalam Audio-text dataset

The macro-average F1-score is 0.40, while the weighted average F1-score is 0.53(refer Table 5)

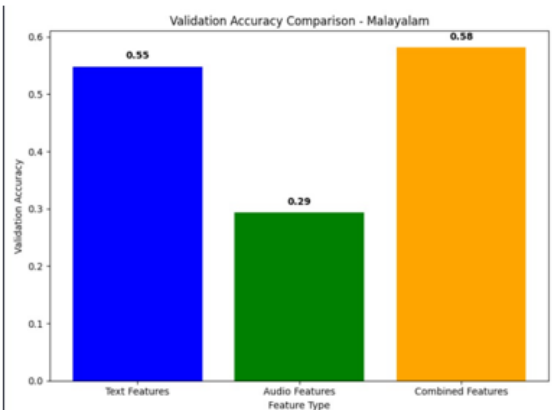


Figure 3: Validation Accuracy Comparison for Malayalam Using Different Feature Types

For Teleugu text data the macro-average F1-score is 0.37, while the weighted average F1-score is 0.51

Category	Precision	Recall	F1-Score	Support
C	0.15	0.16	0.16	25
G	0.18	0.19	0.19	21
N	0.28	0.28	0.28	39
P	0.10	0.08	0.09	12
R	0.21	0.21	0.21	14

Table 6: Accuracy of Telugu audio dataset

The macro-average F1-score is 0.19, while the weighted average F1-score is 0.21 (refer Table 6)

Category	Precision	Recall	F1-Score	Support
C	0.56	0.72	0.63	25
G	0.57	0.72	0.64	39
N	0.55	0.43	0.48	14
P	0.00	0.00	0.00	12
R	0.50	0.33	0.40	21

Table 7: Accuracy of Telugu audio-text dataset

The macro-average F1-score is 0.44, while the weighted average F1-score is 0.53. (refer Table 7)

The model achieved 85% accuracy and a macro F1-score of 0.45, effectively handling class imbalance. SMOTE improved recall for minority classes. The text-only model had a lower macro F1-score (0.68), while the audio-only model scored 0.72, showing their complementary roles. The fused model, integrating both, achieved the highest macro F1-score of 0.76, highlighting the importance of multimodal data for hate speech detection in low-resource languages.

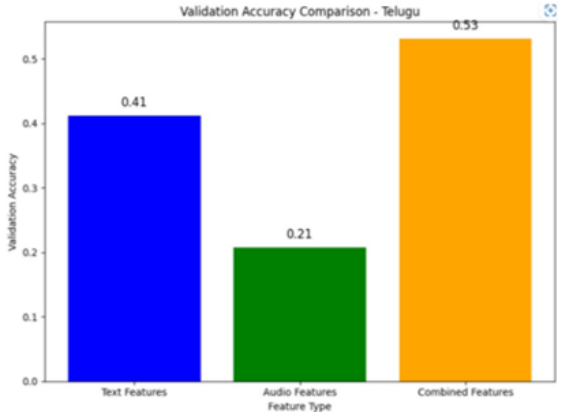


Figure 4: Validation Accuracy Comparison for Telugu Using Different Feature Types

4.1 Limitation

The proposed multimodal hate speech detection system has certain limitations that may affect its performance and scalability. One major limitation is the small and limited dataset, which includes only three Dravidian languages: Tamil, Telugu, and Malayalam. This restricts the model's ability to generalize across diverse linguistic contexts, especially for other lesser-known Dravidian languages. Additionally, the model relies heavily on the availability of both audio and text data, which may not always be practical in real-world scenarios where either of the modalities could be missing or incomplete. Another significant limitation is the class imbalance present in the dataset, where hate speech instances, especially in the "Personal Defamation" category, are underrepresented. Although SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the classes, it may not fully capture real-world data distribution, impacting the model's accuracy. Furthermore, the presence of code-mixed language, where users switch between native languages and English, introduces complexity, as some hate speech expressions may only be detected with a proper understanding of both languages. The model also lacks the inclusion of visual data, such as videos, which could significantly improve hate speech detection by providing context through facial expressions or gestures. Additionally, the study does not assess potential biases in the model, which could impact fairness across different social or cultural groups. Addressing these limitations by expanding the dataset, incorporating video data, and reducing class imbalance could enhance the model's overall performance and inclusiveness.

5 Conclusion

This study highlights effective multimodal hate speech detection in Tamil, Telugu, and Malayalam using Wav2Vec 2.0 and XLM-Roberta for acoustic and textual features. Feature fusion, SMOTE, and XGBoost created a robust system, achieving a macro F1-score of 0.76 and addressing tonal aggression, sarcasm, and contextual nuances. The model's strong performance in underrepresented classes supports future advancements in multilingual hate speech detection.

References

Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Zafar Ali, Sajid Khan, and Ghulam Mujtaba. 2020. Automatic hate speech detection using machine learning: A comparative study. In *Proceedings of the 2020 Conference on Hate Speech Detection*.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. In *Proceedings of the 2022 Conference on Textual Hate Speech Detection*.

Pranjali Prakash Bansod. 2023. Hate speech detection in hindi. In *Proceedings of the 2023 Workshop on Hindi NLP*.

Paula Fortuna and Sérgio Nunes. 2018. Survey on hate speech detection. In *Proceedings of the 2018 International Conference on Hate Speech Detection*.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring hate speech detection in multimodal publications. In *Proceedings of the 2019 Workshop on Multimodal Hate Speech Detection*.

Ahshanul Haque and Naseef Chowdhury. 2023. Hate speech detection in social media using the ensemble learning technique. In *Proceedings of the 2023 Workshop on Ensemble Learning for Hate Speech Detection*.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. In *Proceedings of the 2023 Conference on NLP for Hate Speech*.

Lal G. Jyothish, Premjith B., Chakravarthi Bharathi Raja, Rajiakodi Saranya, B. Bharathi, Natarajan Rajeswari, and Ratnavel Rajalakshmi. 2025. Overview of the shared task on multimodal hate speech detection in dravidian languages: Dravidianlangtech@naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

György Kovács, Pedro Alonso, and Rajkumar Saini. 2021. Challenges of hate speech detection in social media: Data scarcity and leveraging external resources. In *Proceedings of the 2021 Social Media NLP Workshop*.

Peiyu Li. 2021. Achieving hate speech detection in a low resource setting. In *Proceedings of the 2021 Workshop on Low Resource NLP*.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. 2022. Deep learning for hate speech detection: A comparative study. In *Proceedings of the 2022 Conference on Deep Learning for NLP*.

Sara Parker and Derek Ruths. 2023. Is hate speech detection the solution the world wants? In *Proceedings of the 2023 International Conference on Ethical AI*.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. Optimize_prime@dravidianlangtech-acl2022: Abusive comment detection in tamil. In *Proceedings of DravidianLangTech-ACL 2022*.

PK Roy, S Bhawal, and CN Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. In *Proceedings of the DravidianLangTech 2022*.

Taufic Leonardo Sutejo and Dessi Puji Lestari. 2018. Indonesia hate speech detection using deep learning. In *Proceedings of the 2018 Indonesian NLP Conference*.

Amir Toliyat, Sarah Ita Levitan, Zheng Peng, and Ronak Etemadpour. 2022. Asian hate speech detection on twitter during covid-19. In *Proceedings of the 2022 Asian Hate Speech Detection Conference*.

Ching Seh Wu and Unnathi Bhandary. 2020. Detection of hate speech in videos using machine learning. In *Proceedings of the 2020 Workshop on Video-based Hate Speech Detection*.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021. Hate speech detection based on sentiment knowledge sharing. In *Proceedings of the 2021 Conference on Sentiment-based Hate Speech Detection*.

[1]

[1]https://github.com/FarhaAfreen/Multimodal-Hate-Speech-Detection-in-Dravidian-Languages

Wictory@DravidianLangTech 2025: Political Sentiment Analysis of Tamil X(Twitter) Comments using LaBSE and SVM

Nithish Ariyha K, Eshwanth Karti T R,Yeshwanth Balaji AP , Vikash J, Sachin Kumar S
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.u4aie22140@cb.students.amrita.edu,
cb.en.u4aie22118@cb.students.amrita.edu,
cb.en.u4aie22102@cb.students.amrita.edu,
cb.en.u4aie22156@cb.students.amrita.edu, s_sachinkumar@cb.amrita.edu

Abstract

Political sentiment analysis has become an essential area of research in Natural Language Processing (NLP), driven by the rapid rise of social media as a key platform for political discourse. This study focuses on sentiment classification in Tamil political tweets, addressing the linguistic and cultural complexities inherent in low-resource languages. To overcome data scarcity challenges, we develop a system that integrates embeddings with advanced Machine Learning techniques, ensuring effective sentiment categorization. Our approach leverages deep learning-based models and transformer architectures to capture nuanced expressions, contributing to improved sentiment classification. This work enhances NLP methodologies for low-resource languages and provides valuable insights into Tamil political discussions, aiding policymakers and researchers in understanding public sentiment more accurately. Notably, our system secured **Rank 5** in the NAACL shared task, demonstrating its effectiveness in real-world sentiment classification challenges.

Keywords: political sentiment, NLP, SVM, LaBSE, MuRIL, transformer, deep learning, sentiment classification

1 Introduction

The digital age has ushered into society social media, which dramatically changed the discourse of political issues and ushered in totally unprecedented avenues to involve the public within them. Social networking sites, especially platform forums like X are dynamic debating forums that offer a varied mix of thought streams to join a continuous flowing debate. In the digital age, regional languages like Tamil have become indispensable archives of grass-root-level political discourse that unfold genuine insights into local viewpoints and culture-related politically complex expressions. (Gioia, 2023)

A recent area in NLP emerged under the rubric of political sentiment analysis. This aims at providing much-needed insights into the public sentiment by methodical categorization of textual representations of political opinion. Since it supports the measurement of the public's response to policies, measures political engagement, and identifies critical issues in society, this kind of analysis has a high utility for policymakers, political analysts, and governmental agencies. With sentiment analysis, proper policy solutions would be devised more effectively and with alignment to community requirements.

But sentiment classification, per se, is inherently challenging because it often muddles multiple tones: sarcasm, strong opinions, or even seemingly neutral observations. In low-resource languages like Tamil, these challenges are compounded by linguistic and cultural complexities demanding sophisticated techniques for capturing nuances in expression.

This work hopes to address such issues by partitioning Tamil political tweets into seven different groups that support data analysis. In doing so, we look forward to contributing to the further development of NLP methods for low-resource languages while increasing our understanding of Tamil political mood at the same time. This review also aims to provide a comparative analysis of different models and techniques for this task. This work is based on the shared task in DravidanLangTech2025@NAACL (Chakravarthi et al., 2025).

2 Related Works

Political sentiment analysis has gained attention with the rise of social media, particularly X. Review by (Wankhade et al., 2022) discusses about senti-

ment analysis in different areas including social media and e-commerce, methodologies, applications, and challenges. It emphasizes methods like lexicon-based, ML, and hybrid approaches while addressing issues like sarcasm, ambiguity, and language-specific challenges. The study also highlights the impact of emoticons and emojis, which prompted the use of embeddings with ML models for sentiment analysis.

(Elghazaly et al., 2016) compared the Naïve Bayes classifier and SVM classifiers on Arabic tweets in the 2012 Egyptian elections, solving problems like inflectional variation, stemming, and sarcasm. (Babu, 2022) experimented with Tamil sentiment classification on movie reviews using CNN-LSTM, CNN-BiLSTM, and CNN-BiGRU and obtained the maximum accuracy with CNN-BiLSTM. Kumar S (Kumar S et al., 2017) employed CNN and LSTM to classify Malayalam tweets and got better results for identification tasks.

(Tripty et al., 2024) explored ML and DL models for sentiment analysis of YouTube comments, highlighting the strong performance of encoder models like XLM-RoBERTa and IndicBERT. (Kannan et al., 2021) applied IndicBERT to code-mixed Tamil tweets, achieving a 61.73 F1-score, which aligns with our task of classifying political sentiments in Tanglish data.

(Tripty et al., 2024) explored a variety of ML and DL based models for sentiment analysis of youtube comments. Their review revealed how encoder models like XLM-RoBERTa and IndicBERT perform well in the classification task, paving way for model selection in our work. The authors of (Kannan et al., 2021) apply Indic-BERT to analyze code-mixed Tamil tweets and demonstrates its effectiveness over traditional methods. The study achieved an F1 score of 61.73, and this aligns closely with the task of classification of political sentiments for Tanglish data (Tamil and English).

(Kumar and Albuquerque, 2021)'s study shows the performance of XLM-R large model in comparison with models like BB_Twtr and DataStories . The XLM-R large model surpasses the rest models by 5% with 71.8% accuracy. Authors of (Nithya et al., 2022) aim to apply deep learning based Bi-LSTM model with ULMFiT for sentiment analysis, which gave them promising and better results. Authors of (Shanmugavadivel et al., 2022) provide and analysis of multiple machine learning models for sentiment analysis of Tamil code-mixed data. They tested many methods like SVM, Logistic Regressions, CNN, BiLSTM and many with their best F1 score being around 0.66.

3 Dataset

The dataset used in this study was obtained from the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics. The dataset is pre-divided into training and test subsets, with the test set comprising approximately 430 sentences. The training dataset consists of about 4,300 Tamil sentences, where each sentence contains both hashtags and emojis, categorized into seven distinct classes: Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral, and None of the Above. The lengths of the sentences are predominantly between 100 and 250 characters. With the dataset size being small, it is also imbalanced. It contains more samples in Opinionated than in the None of the Above category.

4 Methodology

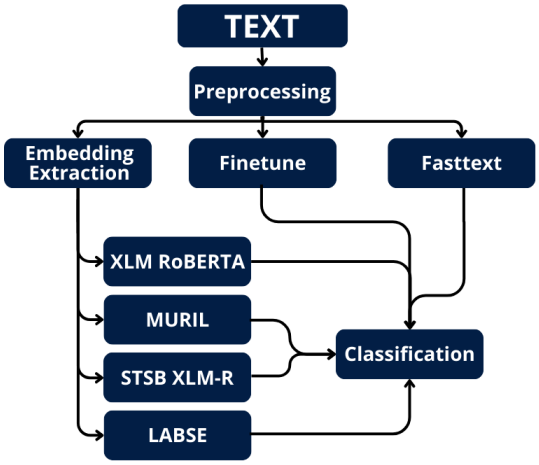


Figure 1: Experimentation pipeline

Fig. 1 shows the experimentation pipeline that was followed.

4.1 Data Preprocessing

The preprocessing techniques of this study center on cleaning and normalizing textual data for better quality and reliability of downstream NLP tasks. First, emojis are replaced by their text equivalent to ensure contextual information. Hashtags are replaced with the tag <*hashtag*> for maintaining tweet data. As the emojis contain meaningful sentiment-related information, simply deleting or replacing with a generic tag as done with hashtags

would compromise model performance. All text is further converted to lowercase, removing special characters, extra spaces (including tabs and newlines), URLs, and mentions to improve uniformity. Other words with certain symbols like currency signs or special characters are also removed, as they could interfere in unwanted ways during text analysis. Overall, all these preprocessing steps lead to noise reduction, text normalization, and an overall improvement in the quality of data. Finally, labels and content are formatted according to requirements, such as converting labels to numerical values or adapting them to FastText format.

To improve the class imbalance we tried the method of ADASYN(Adaptive Synthetic Sampling). ADASYN is an oversampling technique that generates synthetic samples for the minority class based on data distribution (He et al., 2008). This method was used only in one trial due to the constraints of time and compute available.

4.2 Fine-Tuning Transformer models

Transformer-based models such as IndicBERT, MuRIL, TamilSBERT (Joshi et al., 2022), XLM-RoBERTa, and TamilBERT4MLM were trained with Adam optimizer and weight decay, batch size 8, and sequence length of 256(chose based on text distribution). Focal Loss was used to address class imbalance, while exploding gradients were avoided with gradient clipping . FastText's skip-gram model employed 300-dimensional vectors and a subword n-gram of three to six characters to adapt to the morphological richness of Tamil and code-mixed tokens (Bojanowski et al., 2017). F1-score was the primary evaluation metric.

4.3 Embedding with SVM Classifier

To effectively classify text using Support Vector Machines (SVM), we first extracted meaningful embeddings from multiple transformer-based and static embedding models. The embeddings served as input features to the SVM classifier, ensuring that the model had a well-represented feature space for learning. Below, we provide details on each stage of this process.

4.3.1 Embedding Extraction

To generate high-quality embeddings from each model, we employed distinct extraction strategies tailored to their respective architectures.

XLM-RoBERTa: Mean pooling over the last hidden state to obtain sentence representations that capture linguistic structures effectively. (Conneau et al., 2019)

LaBSE: Sentence representations were extracted and normalized to ensure uniform magnitude across different inputs, enhancing stability in classification. (Feng et al., 2022)

MuRIL: The classification token $[CLS]$ representation was used, leveraging its pre-training to encode sentence-level semantics efficiently.

STSb-XLM-R: Token embeddings were mean-pooled to generate comprehensive sequence representations.

4.3.2 ML Based Classification

For classification, we employed a Support Vector Machine (SVM) approach, using Scikit-learn's LinearSVC implementation to ensure computational efficiency and reliable performance.

To handle class imbalance, class weights were set inversely proportional to class frequencies. This strategy prevented minority classes from being underrepresented during training, ensuring that the classifier made balanced predictions across all categories.

Features were normalized employing Robust Scaler. Grid search tuned the regularization parameter $\{0.1, 1, 10\}$ to harmonize margin maximization and performance in classification. Combining structured embeddings with a classifier based on SVM efficiently made sentiment classification operational in Tamil-English code-mixed political utterances.

After the shared task closure, XGBoost was also experimented. It builds decision tree sequentially, correcting previous errors while minimizing loss. With L1/L2 regularization if handles overfitting. It is efficient in handling missing values, and parallel processingmakes it is highly scalable. LABSE embeddings trained the model using TF-IDF, enriching the input with more information.

5 Results and Observations

All transformer models had good accuracy on the initial stages of training, but the accuracy started to saturate at around 30%, indicating a trade-off limitation between overfitting and bad generalization beyond some point. With increasing overfitting to the training data, which is seen to grow large with respect to the gap between training and validation performance, the tested models produced different performances. TamilSBERT performed the best at an accuracy of 38% followed by F1 Score 0.26,

Table 1: Embedding Models and their Performance

Model	Weighted F1
FastText	0.280*
XLM-RoBERTa Base + SVM	0.220
Sentence-Transformers-LaBSE + SVM	**0.310***
MuRIL + SVM	0.150
Indic-Bert + SVM	0.24
STSb-XLM-R	0.260
Sentence-Transformers-LaBSE + XG-Boost + TFIDF + ADASYN	**0.330***

Weighted F1 Score

showing slight improvement over the others. The highest accuracy in TamilSBERT was due to a better tokenizer as it could store all word data while other tokenizers could not. The overfitting problem continued with all the models, showing that there is a need for better and diversified training data for better generalization of transformer-based architectures for Tamil English code-mixed political discourse.

From the outcome as shown in Table 1, the Sentence-Transformers-LaBSE model along with the SVM classifier performs best in SVM, reporting the highest value with an accuracy of 0.310 in the case of a weighted F1 score. This shows, LaBSE embeddings specifically optimized for sentence-level multilingual tasks are able to handle the text in Tamil English code-mixed variety, especially in low-resource settings. Post Shared task experiments showed XGBoost with LABSE and TF-IDF displayed the highest F1 score of 0.33

FastText performed well with a weighted F1 score of 0.280, showing its effectiveness in handling morphologically rich languages like Tamil. Although it is a multilingual model, XLM-RoBERTa Base scored a lower score at 0.220, possibly because it has not been exposed to much Tamil English code-mixed data. STSb-XLM-R showed moderate performance with an F1 score of 0.260, showing its ability to capture contextual relationships. Surprisingly, MuRIL, designed for Indian languages, had the lowest score (0.150), suggesting its pretraining may not sufficiently cover Tamil English code-mixed text. On looking close into the classification scores for each class a common pattern was observed. The class of none of the above showed a significantly high score of 0.79 whereas the class substantiated achieved only 0.11. The primary reason for this being the similarity in substantiated and opinionated. This could be the primary reason for average F1 scores.

The very low F1 values for Tamil-English code-mixed political opinion analysis can be attributed to a variety of reasons. The primary reason is the lack of sufficient and quality labeled data, which affects the ability of the model to learn informative patterns. A improper train-test split, with the test set including entirely unseen tokens, will also prevent generalization. The intricacy of code-mixed language, such as variations in grammar, inconsistency in transliteration, and varying word orders, makes it even more challenging. Most pretrained language models, also, are not specially trained for Tamil-English code-mixed data, which restricts their performance. A potential future improvement is using large language models to address these issues effectively.

Even though there are numerous research and reviews on sentiment analysis of Tamil-English code-mixed text, we couldn't compare our results with them due to their simple classification approach. The majority of work in this area focuses on basic sentiment analysis, such as classifying text as positive or negative, rather than a more detailed classification. This highlights a greater scope for future research in this area.

6 Conclusion

This report presents the findings from the sentiment analysis task conducted as part of the FifthWorkshop on Speech and Language Technologies .The task focused on classifying political sentiments in Tamil English code-mixed tweets, with the dataset provided by the conference. Our proposed method achieved a rank of 5th in the overall task. The results demonstrate the effectiveness of leveraging both transformer-based models and traditional embeddings for sentiment classification in low-resource languages, while highlighting the need for further improvements in handling code-mixed text for better generalization and better datasets.

Link for GitHub repository with codes[1]

[1] https://github.com/ariyha/NAACL-2025-Political-Sentiment-Analysis

7 Limitations

The volume of the given dataset could be expanded to capture greater variability, ensuring that deep learning models are trained on a more diverse representation of political discourse. Additionally, political tweets often include multimodal elements such as images, videos, memes, and emojis, which are not accounted for in text-only sentiment analysis models, potentially leading to incomplete or inaccurate sentiment predictions. Another crucial limitation is the issue of concept drift, where models trained on past data may become outdated as political narratives evolve over time. Therefore, sentiment models should not be static; they must be continuously updated to adapt to shifts in public opinion and emerging political contexts. This ongoing evolution is essential for real-world applications, where accurate sentiment analysis depends on the model's ability to reflect current sociopolitical dynamics rather than relying solely on historical data.

References

Suba Sri Ramesh Babu. 2022. Sentiment analysis in tamil language using hybrid deep learning approach. Msc research project, National College of Ireland.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Preprint*, arXiv:1607.04606.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Tarek Elghazaly, Amal Mahmoud, and Hesham A. Hefny. 2016. Political sentiment analysis using twitter data. In *Proceedings of the International Conference on Internet of Things and Cloud Computing*, ICC '16, New York, NY, USA. Association for Computing Machinery.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Elio Simone La Gioia. 2023. Using sentiment analysis for politics: the case of the italian political elections.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

Ananya Joshi, Aditi Kajale, Janhavi Gadre, Samruddhi Deode, and Raviraj Joshi. 2022. L3cube-mahasbert and hindsbert: Sentence bert models and benchmarking bert sentence representations for hindi and marathi. *arXiv preprint arXiv:2211.11187*.

R. Ramesh Kannan, Ratnavel Rajalakshmi, and Lokesh Kumar. 2021. Indicbert based approach for sentiment analysis on code-mixed tamil tweets. In *Fire*.

Akshi Kumar and Victor Hugo C Albuquerque. 2021. Sentiment analysis using xlm-r transformer and zero-shot transfer learning on resource-poor indian language. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–13.

Sachin Kumar S, Anand Kumar Madasamy, and Soman Kp. 2017. *Sentiment Analysis of Tweets in Malayalam Using Long Short-Term Memory Units and Convolutional Neural Nets*, pages 320–334.

K. Nithya, S. Sathyapriya, M. Sulochana, S. Thaarini, and C. R. Dhivyaa. 2022. Deep learning based analysis on code-mixed tamil text for sentiment classification with pre-trained ulmfit. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1112–1116.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech Language*, 76:101407.

Zannatul Tripty, Md. Nafis, Antu Chowdhury, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUETSentimentSillies@DravidianLangTech-EACL2024: Transformer-based approach for sentiment analysis in Tamil and Tulu code-mixed texts. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 234–239, St. Julian's, Malta. Association for Computational Linguistics.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7a):5731–5780.

ANSR@DravidianLangTech 2025: Detection of Abusive Tamil and Malayalam Text Targeting Women on Social Media using RoBERTa and XGBoost

Nishanth S, Shruthi Rengarajan, S. Ananthasivan, Burugu Rahul, Sachin Kumar S
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
{cb.en.u4aie22149, cb.en.u4aie22154,
cb.en.u4aie22148, cb.en.u4aie22161}@cb.students.amrita.edu
s_sachinkumar@cb.amrita.edu

Abstract

Abusive language directed at women on social media, often characterized by crude slang, offensive terms, and profanity, is not just harmful communication but also acts as a tool for serious and widespread cyber violence. It is imperative that this pressing issue be addressed in order to establish safer online spaces and provide efficient methods for detecting and minimising this kind of abuse. However, the intentional masking of abusive language, especially in regional languages like Tamil and Malayalam, presents significant obstacles, making detection and prevention more difficult. The system created effectively identifies abusive sentences using supervised machine learning techniques based on RoBERTa embeddings. The method aims to contribute to a safer cyber space from abusive language, which is essential for various online platforms -including but not limited to- social media, online gaming services etc. The proposed method has been ranked **8** in Malayalam and **20** in Tamil in terms of *f1* score.

Keywords: Abusive texts, Social Media, Natural Language Processing, RoBERTa, XGBoost

1 Introduction

Social media are found to be the new entertainments, information mediums, and communications alike. However, they also present online harassments by targeting women. The negative consequences such content has for victims have serious psychological, social, and professional impacts that highlight the need for effective tools to detect and mitigate abuse. Abuse appears as hate, abusive, or threatening comments directed at others as deep-rooted societal biases with the intent to promote gender inequality. Therefore, mechanisms are needed for the identification and mitigation of online abuse.

Although many work has been reported on abusive language detection for high-resource languages such as English, little work has been done on low-resource languages (Chakravarthi et al., 2023). Here, two popular South Indian languages- Tamil (Rajalakshmi et al., 2023) and Malayalam are taken(Raphel et al., 2023) -with minimal annotated datasets and tools for Natural Language Processing (NLP). Detecting abusive language is made harder due to various linguistic complexities, code mixing, and dialectal variations; therefore, this becomes an area of concern.

To address this challenge, our team participated in the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media at DravidianLangTech@NAACL 2025 (Rajiakodi et al., 2025). More details about the shared task can be found at[1].

We used Machine Learning and Natural Language Processing techniques were used to build an automated detection system for abusive content directed at women (Hossain et al., 2022). Our approach is by using pre-trained word embeddings and fine-tuning an XGBoost model to achieve effective classification. We are, therefore, trying to contribute towards safer digital environments and support the efforts of moderation of content in social media with the development of Machine learning models for these low-resource languages.

2 Dataset

The data set was distributed by the shared task organisers of 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (Priyadharshini et al., 2022, 2023).

The dataset used for this study comprises sentences in Tamil and Malayalam languages, categorized into two classes: *Abusive* and *Non-Abusive* (Class distribution). The data set is split into train-

[1] https://codalab.lisn.upsaclay.fr/competitions/20701

Data Type	Total Sentences	Class Distribution
Training	3562	1728 : 1834
Testing	629	303 : 326

Table 1: Dataset Statistics for Malayalam

Data Type	Total Sentences	Class Distribution
Training	3388	1644 : 1744
Testing	598	293 : 305

Table 2: Dataset Statistics for Tamil

ing and test sets. Table.1 and Table.2 show the data distribution of the Malayalam and Tamil classes

It was specifically developed for evaluating the suitability of language models for identifying abusive language in low-resource Dravidian languages, ensuring near-balanced *Abusive* and *Non-Abusive* example representations to provide more efficient training and evaluation.

3 Methodology

The flowchart given in Figure.1 describes the proposed methodology used in the classification of the abusive and non-abusive comments of both Tamil and Malayalam (Kumar et al., 2017).

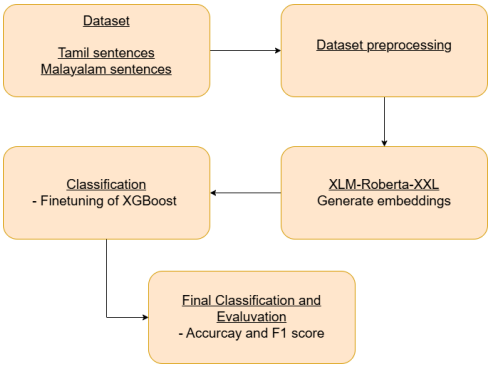


Figure 1: The proposed methodology

3.1 Data Preprocessing

The data preprocessing workflow starts with every dataset going through text cleaning, where we strip URLs, special characters, and unnecessary symbols and convert all text to lowercase to ensure uniformity and to improve model. Additionally, the target class labels (e.g., *abusive* and *non-abusive*) are standardized to lowercase for consistency.

After cleaning, the labels are mapped to numerical values: *abusive* is assigned the value `1`, and *non-abusive* is assigned the value `0`. This numerical encoding is essential for machine learning algorithms, which require numeric inputs for supervised learning tasks. Finally, the training and development datasets are concatenated into a combined dataset to ensure that the model is trained on a more diverse and comprehensive set of examples. Both Tamil and Malayalam datasets are combined as only a single model is developed for this classification task. The train-test split 80-20.

3.2 Feature Extraction and Model Preparation

XLM-RoBERTa-XXL model (Goyal et al., 2021), a state-of-the-art transformer-based model (Vaswani et al., 2017) pre-trained on multilingual corpus was used to get word embeddings for the model to be trained on.It is particularly suited for handling low-resource languages like Tamil and Malayalam. The tokenizer's job is to encode the text data into input features suitable for the transformer model, while the configuration ensures that the model's architecture and hyper-parameters align with the expected outcome, i.e., to classify abusive content. The model configuration details are shown in Table.3.

Property	Details
Parameters	10.7 billion
Number of Layers	48
Embedding Dimensions	4096

Table 3: XLM-RoBERTa-XXL Model Details

3.3 Embedding Generation

Extraction embeddings use the XLM-RoBERTa-XXL model and tokenizer to generate high dimensional vector representations of text data.

For each text sequence, the tokenizer encodes the input by truncating it to a specified maximum length of 512 tokens. The tokens are then passed through the model, and the output embeddings are obtained. A function then extracts the embedding of the first token (CLS token) from the model's output, which represents a summary of the entire sequence. These embeddings are then stored and concatenated into a single tensor, giving the final vector representation for the input data.

3.4 Machine Learning Models

Various machine learning models like logistic regression, K-Nearest Neighbors (KNN) and random forest, were explored (V P et al., 2023; Hasan et al.,

2024; K et al., 2021). The scores of these models are shown in Table.4.

Method	F1 Score
Logistic Regression	0.6618
K-Nearest Neighbors	0.5832
Random Forest	0.6735

Table 4: Scores from Different Models

After carefully testing the performance of each of the different models, XGBoost (Chen and Guestrin, 2016) stood out for the following reasons:

- **Handling Complex Relationships:** Unlike models like logistic regression and Random Forest, XGBoost captures complex patterns and interactions in the data, which is essential when working with high-dimensional word embeddings.
- **Efficiency:** XGBoost is optimised for speed and performance, making it faster to train and more efficient than models like random forest and KNN, especially while training on larger datasets like this.
- **Flexibility in Tuning:** XGBoost has a lot of hyperparameters that can be adjusted to gain better performance, including learning rate, maximum tree depth, and regularization terms.
- **Regularization:** XGBoost has built-in L1 and L2 regularization that helps to prevent overfitting, which is a major cause for concern with other models like SVM when using large embeddings.

These advantages made XGBoost an easy choice for building the pipeline.

3.5 Model Training and Submission

During the initial submission, a single XGBoost model was trained to handle both Tamil and Malayalam text simultaneously. Instead of training separate models for each language, this two birds in one shot approach shared patterns across the two languages that reduced the computational requirements (Koreddi et al., 2025).

3.6 Performance Metrics

The results from this submission was ranked as as shown in Table.5.

Language	Rank	Macro F1 Score
Malayalam	8th	0.6901
Tamil	20th	0.7201

Table 5: Submission Results in Malayalam and Tamil

4 Experimental Results

In order to determine which combination of hyperparameters produced the best outcomes, approximately 324 (36×9, different $n_{\text{estimators}}$) setups were executed. The plotted graph is shown in Figure.2. According to this analysis, the top three configurations were chosen for subsequent testing.

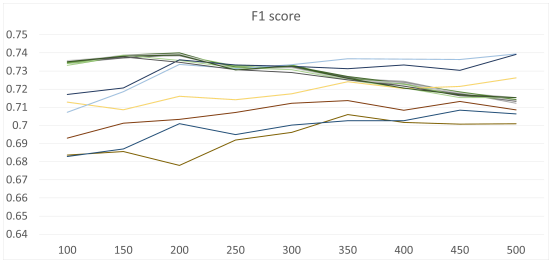


Figure 2: Accuracy Trends Across Experiments

These settings were subsequently tested using a further 30 (3×10, different $n_{\text{estimators}}$) settings to optimize the value of $n_{\text{estimators}}$ for optimal performance without sacrificing computational efficiency. The model configurations are shown in Table.6.

Hyperparameter	Value
Objective	binary:logistic
Max Depth	6
Learning Rate	0.1
Random State	42
Tree Method	hist
Device	cuda
Number of Estimators	1000
Evaluation Metric	error
Booster	dart
Subsample	0.5

Table 6: XGBoost Model Configuration after finetuning

This iterative process helped to find the highest-performing setup, significantly better than the initial baseline, while maintaining the computational speed for which XGBoost is renowned. The best F1 scores given by the model is shown in Figure.3.

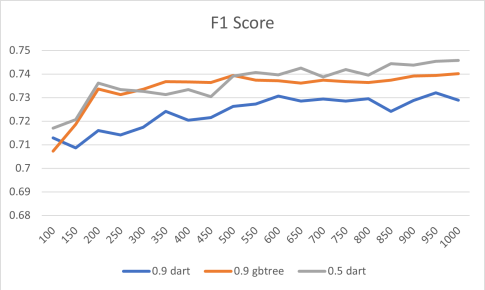


Figure 3: F1 Score of best 3 configurations

5 Evaluation

The performance of the fine-tuned XGBoost model was measured by employing the F1 score as the major indicator. The model was tested and trained with other methodologies to observe how well it would perform when processing Tamil and Malayalam text.

5.1 Comparison with Other Approaches

Prior to finalizing RoBERTa-based embeddings, extraction of embeddings from IndicBERT was attempted. When trained with XGBoost, the IndicBERT embeddings gave an F1 score in the mid-60s (~ 0.65). From this, we had a goal of reaching an F1 score in the 70s by trying more efficient embedding methods.

Shifting towards RoBERTa-based embeddings (XLM-RoBERTa), we saw significant improvement, resulting in an F1 score of **0.71** upon training and validation both on Tamil and Malayalam simultaneously. This attested that RoBERTa embeddings performed better to identify linguistic aspects pertaining to abusive language detection and was subsequently employed for initial submission.

5.2 Final Model Performance

With the optimized XGBoost model based on RoBERTa embeddings, the optimal F1 score obtained is as shown in Table.7.

Data	F1-Score
Tamil and Malayalam combined	0.745
Tamil only	0.775
Malayalam only	0.713

Table 7: Final Output Evaluation: F1 Scores Breakdown

These findings show that although a common model for both languages works quite well, training on each language separately gives a minor performance improvement. This implies that language-specific subtleties may have an effect on classification performance, and additional optimizations like language-aware pre-processing or more feature engineering may improve results (Shubhankar Barman, 2023).

The code files for this project can be accessed from[2]

6 Conclusion

This paper presented the results of the task performed as part of the Fifth Workshop on Speech and Language Technologies for Dravidian Languages on abusive text detection in Tamil and Malayalam dataset on women in social media. The conference provided the dataset for the proposed task. Out of 156 participation and 30 submissions, this proposed method was ranked 8 in Malayalam dataset and 20 in Tamil dataset.

7 Limitations

Despite getting promising results, the methodology has certain limitations that could affect performance and scalability:

- **Large Model Size:** The use of RoBERTa-XXL embeddings significantly increased the computational requirements. Due to the model's large size, standard GPUs were insufficient, and an `NVIDIA A6000` was required to handle the memory load. This makes the approach less accessible for environments with limited hardware resources.

- **High Computational Costs:** Training and fine-tuning models on such large embeddings required extensive computational time and resources, which may not be feasible for all researchers or in production environments.

References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting Abusive Comments at a Fine-Grained Level in a Low-Resource Language. *Natural Language Processing Journal*, 3:100006.

[2] https://github.com/ANSR-codes/NAACL_Shared_task

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-Scale Transformers for Multilingual Masked Language Modeling. *CoRR*, abs/2105.00572.

MD. Nahid Hasan, Kazi Shadman Sakib, Taghrid Tahani Preeti, Jeza Allohibi, Abdulmajeed Atiah Alharbi, and Jia Uddin. 2024. OLF-ML: An Offensive Language Framework for Detection, Categorization, and Offense Target Identification Using Text Processing and Machine Learning Algorithms. *Mathematics*, 12(13).

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, M. Ali Akber Dewan, Nazmul Siddique, and Md. Azad Hossain. 2022. Identification of Multilingual Offense and Troll from Social Media Memes Using Weighted Ensemble of Multimodal Features. *Journal of King Saud University - Computer and Information Sciences*, 34(9):6605–6623.

Sreelakshmi K, Premjith B, and Soman Kp. 2021. Amrita_CEN_NLP@DravidianLangTech-EACL2021: Deep learning-based offensive language identification in Malayalam, Tamil and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 249–254, Kyiv. Association for Computational Linguistics.

Venkatesh Koreddi, Nalluri Manisha, Shaik Kaif, and Yeligeti Kumar. 2025. Multilingual AI System for Detecting Offensive Content Across Text, Audio, and Visual Media.

S. Sachin Kumar, M. Anand Kumar, and K. P. Soman. 2017. Sentiment Analysis of Tweets in Malayalam Using Long Short-Term Memory Units and Convolutional Neural Nets. In *Mining Intelligence and Knowledge Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings*, page 320–334, Berlin, Heidelberg. Springer-Verlag.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Faerie Mattins R., Pavitra Vasudevan, and Anand Kumar M. 2023. Hate and Offensive Content Identification in Tamil Using Transformers and Enhanced Stemming. *Computer Speech Language*, 78:101464.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Mariya Raphel, Premjith B, Sreelakshmi K, and Bharathi Raja Chakravarthi. 2023. Hate and Offensive Keyword Extraction from CodeMix Malayalam Social Media Text Using Contextual Embedding. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 10–18, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mithun Das Shubhankar Barman. 2023. Multimodal Abusive Language Detection and Sentiment Analysis in Dravidian Languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*.

Abeera V P, Dr. Sachin Kumar, and Dr. Soman K P. 2023. Social media data analysis for Malayalam YouTube comments: Sentiment analysis and emotion detection using ML and DL models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762.

Synapse@DravidianLangTech 2025: Multiclass Political Sentiment Analysis in Tamil X (Twitter) Comments: Leveraging Feature Fusion of IndicBERTv2 and Lexical Representations

Suriya KP[1], **Durai Singh K**[1], **Vishal A S**[1], **Kishor S**[1], **Sachin Kumar S**[1]
[1]Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
{cb.en.u4aie22164, cb.en.u4aie22167, cb.en.u4aie22159, cb.en.u4aie22128}
@cb.students.amrita.edu , s_sachinkumar@cb.amrita.edu

Abstract

Social media platforms like X (Twitter) have gained popularity for political debates and election campaigns in the last decade. This creates the need to moderate and understand the sentiments of the tweets in order to understand the state of digital campaigns. This paper focuses on political sentiment classification of Tamil X (Twitter) comments which proves to be challenging because of the presence of informal expressions, code-switching, and limited annotated datasets. This study focuses on categorizing them into seven classes: substantiated, sarcastic, opinionated, positive, negative, neutral, and none of the above. This paper proposes a solution to Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments - DravidianLangTech@NAACL 2025 shared task, the solution incorporates IndicBERTv2-MLM-Back-Translation model and TF-IDF vectors into a custom model. Further we explore the use of preprocessing techniques to enrich hashtags and emojis with their context. Our approach achieved Rank 1 with a macro F1 average of 0.38 in the shared task.

Keywords: Political Comments, Tamil tweets, NLP, IndicBERTv2, TF-IDF, Agentic System, Sentiment Analysis.

1 Introduction

Sentiment analysis is a method of analyzing and interpreting feelings, attitudes, and opinions in text. Aspects of sentiment analysis are an integral part of surveys on public opinion, election-future predictions, and policymaking in terms of political discourse. Public forums such as X (Twitter), have emerged as an active platform where people express their opinions in real time, every day, providing a rich resource for analyses of this type (V P et al., 2023). The spontaneity and openness of social media make it a treasure trove to study public sentiments, especially in diverse lingual settings.

Tamil is a Dravidian language, which is rich in literary traditions and predominantly spoken in Tamil Nadu, India, and some parts of Sri Lanka, Singapore and Malaysia. However, culturally important, Tamil is a low-resource language in terms of NLP, lacking annotated datasets, computational tools, and language-specific resources. That marks significant challenges in Tamil sentiment analysis. The informal nature of online speech, slang usage, code-mixing with English (Sreelakshmi et al., 2024), and the frequent use of emojis, hashtags, and abbreviations demand careful linguistic interpretation to convey its complexity.

To tackle these challenges, the shared task "DravidianLangTech@NAACL 2025" (Chakravarthi et al., 2025) will focus on political multiclass sentiment analysis of Tamil X (Twitter) comments to analyse Tamil political discourse while being aware of its linguistic complexities. This paper explores the use of IndicBERTv2, which is trained in 23 Indian languages, for deep contextual understanding and TF-IDF for lexical feature extraction, thus improving the robustness of the model.

2 Related Work

In recent years, significant research has focused on political sentiment analysis, particularly in identifying and understanding public opinion and bias within social media. (B et al., 2024) employed TF-IDF with n-gram features in ensemble models for Tamil and Tulu sentiment analysis achieving F1 scores of 0.260 and 0.550, respectively. The authors demonstrated the capability of TF-IDF in feature extraction for code-mixed tweets providing deep learning representations.

(Kumar et al., 2017) employed a BiLSTM-CNN model for sentiment analysis in Malayalam, attaining an accuracy of 0.9824 has allowed more room for the introduction of transformers in low-resource languages. (Chakravarthi et al., 2021) obtained the

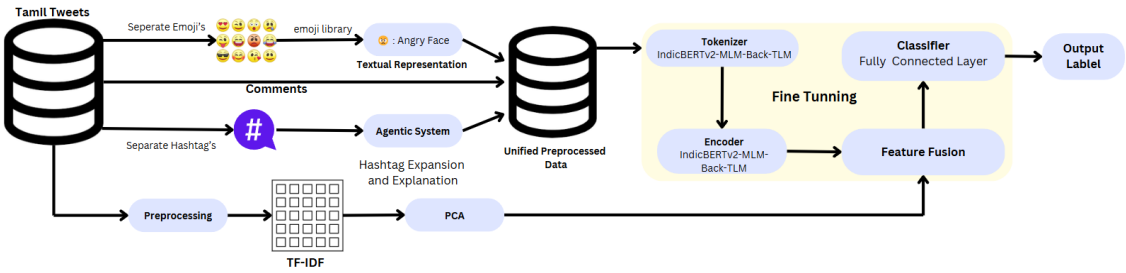


Figure 1: proposed workflow for Political Sentiment Analysis in Tamil X (Twitter) comments. The workflow comprising the modules of preprocessing, feature extraction, model training, and classification.

weighted F1 score of 0.711 in the Tamil-English sentiment analysis based on political psycholinguistic studies in multilingual settings. (Rajalakshmi et al., 2022) combined MuRIL with emoji-based sentiment analysis and proved that the introduction of emojis improved classification accuracy and said that multimodal features are essential in code-mixed sentiment analysis. Finally, (Angdresey et al., 2025) proposed a hybrid model based on tagging along with a BERT-based model, random oversampling, and Multinomial Naïve Bayes, achieving an accuracy of 85.155% and AUC of 96.80% in political sentiment analysis of YouTube comments.

3 Dataset

The dataset has been annotated for political sentiments in Tamil X (Twitter) comments and included comments, which were annotated into seven categories: Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral, and None of the above.

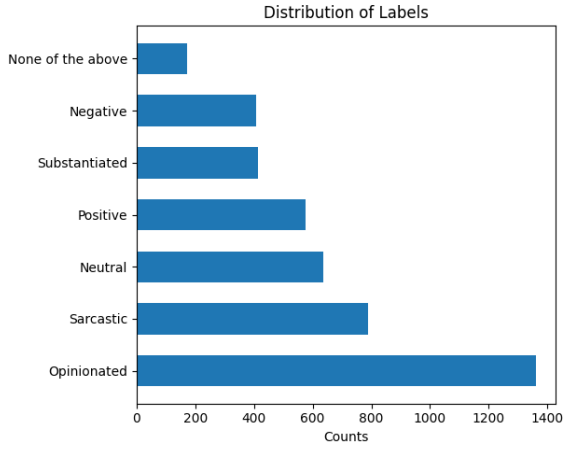


Figure 2: Class distribution in the training dataset.

In total, the dataset has 5440 comments which is divided into three subsets: 80% samples for training with a count of 4,352 samples, 10% for development with 544 samples, and 10% for testing with 544 samples. As organized, the dataset is pre-stratified by the organizers for the sake of uniformity among splits, but Figure 2 depicts the imbalance distribution of different classes in training data.

To balance this, class weights were calculated as the inverse of the class frequencies. These weights were then added to the model's training process to ensure a more balanced learning approach, allowing the model to handle underrepresented categories effectively.

4 Methodology

4.1 Pre-processing

Our pre-processing pipeline has been designed considering the informal and code-mixed nature of Tamil political tweets to ensure that there is meaningful text representation for sentiment classification. The following are the steps it contains:

Demojize: Emojis are converted into textual descriptions using the python emoji library. This step helps retain the emotions expressed through emojis, improving sentiment classification accuracy.

Agentic System: The system uses a two-step process: context retrieval and contextual generation. The most commonly used 160 hashtags were filtered, taking relevant hashtags. For each hashtag, the system queries the Serper API for three top-ranked results and captures contextually relevant text snippets to be used as context. Then, with the help of LLaMA 3.1, it forms a short one-line description depending on pre-existing rules, such as Tamil abbreviations being expended, political actors being identified, and parties being tagged. Such a method applies precise, context-sensitive hashtag explanations, enhancing the sentiment analysis for Tamil political debates.

Stopword Removal: Commonly occurring but non-informative words are removed to improve the

effectiveness of feature extraction. This will be used for Term Frequency-Inverse Document Frequency (TF-IDF) representation, as it focuses on meaningful words that contribute to sentiment analysis.

4.2 Feature Extraction

After pre-processing, feature extraction techniques are employed to convert the text into some numerical representation suitable for training. The main techniques used here are:

Term Frequency-Inverse Document Frequency: TF-IDF assigns importance to the words in a document based on their frequency while reducing the weight of frequent words that appear in many documents. It improves feature representation in text classification by highlighting distinctive terms. Higher values indicate more informative words.

$$\text{TF-IDF}(txt, doc) = \text{TF}(txt, doc) \times \log\left(\frac{N'}{\text{DF}(txt)}\right) \quad (1)$$

Where N' denotes total number of document. and $\text{DF}(txt)$ denotes number of documents that contain the term txt.

TF-IDF was employed to enrich IndicBERTv2 by emphasizing important words that transformers may not catch, enhancing interpretability, capturing domain-specific phrases

Principal Component Analysis(PCA):Reduces the dimension of the data while preserving essential variance, making it useful for noise reduction and visualization.

$$X' = XW \quad (2)$$

where W consists of eigenvectors of the covariance matrix of X, capturing the most significant variance in the dataset.

PCA was applied for dimensionality reduction, noise removal, and preserving vital variance to ensure efficient and concise feature representation

5 Model and Training

In our model, we propose a hybrid approach, combining deep contextual embeddings from **IndicBERTv2-MLM-Back-Translation** (Doddapaneni et al., 2023) with features obtained from **TF-IDF** vectorization (S N et al., 2022), allowing the model to exploit both semantic knowledge and frequency-based linguistic patterns to enhance classification performance. The entire architecture of the model is shown in the figure 3.

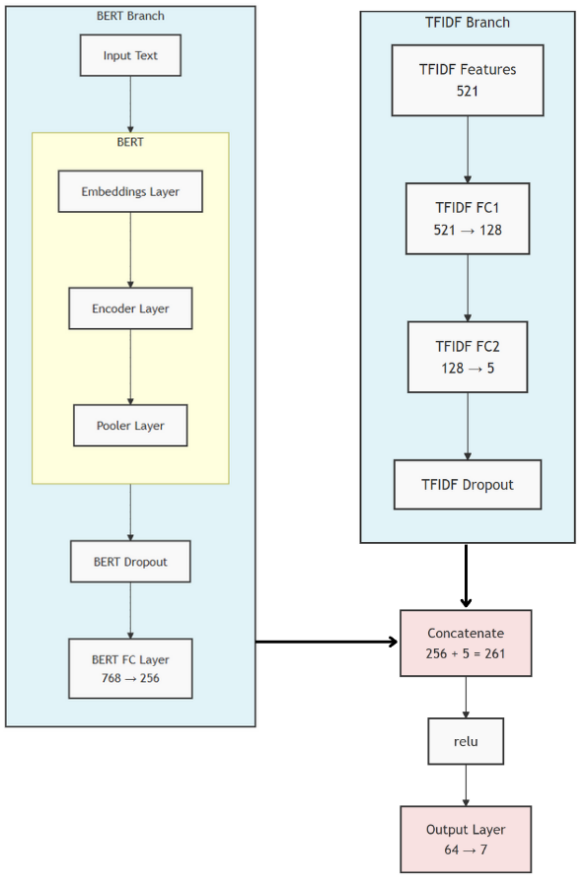


Figure 3: Architecture of the IndicBERTv2-based classification model used in the competition.

There are two main components of the system: a transformer-based encoder and a statistical feature-based discriminator. Preprocessed comments are passed to IndicBERTv2 model to process input text and extract contextual features which get concatenated with the TF-IDF features. These features are then passed through a fully connected classification network with dropout regularization to control overfitting. Finally, a softmax layer predicts one from among the seven sentiment categories.

The training setup has several optimization steps that aim to improve model generalization. We have used the **AdamW** optimizer with the learning rate set at 2e-5 with **weighted CrossEntropyLoss**, taking into account class imbalance by weighing according to inverse class frequencies. To have stable learning, we used **incremental batch sizes**. That is, the batch size is initialized to 16 for the first two epochs and then increased to 32, 48, and finally 64 for the later epochs. Seeded shuffling of dataset is

performed before every epoch to maintain variance in the distribution of classes throughout all passes.

The model is fine-tuned for eight epochs, allowing both representations to move into more nuanced representations of Tamil's political discourse. To control overfitting, dropout regularization is applied with a regularization probability of 0.1 for both models' feature sets, i.e., the transformer and TF-IDF.

6 Result And Analysis

6.1 Macro Average F1-Score

We have used the F1 macro average, as required in the task, which computes the harmonic mean of precision and recall for each class and then averages across all classes, which is appropriate for handling imbalanced datasets by giving equal treatment to all classes.

$$F1_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \times P_i \times R_i}{P_i + R_i} \quad (3)$$

Where C denotes Total number of classes and R_i and P_i denotes precision and recall of class i respectively.

This metric ensures fair performance evaluation, even when some classes are underrepresented.

6.2 Results

We conducted experiment on four different IndicBERT-V2 models as our base encoder, the results are provided in Table 1

Model	Macro-F1
IndicBERTv2-MLM-Back-TLM	0.383
IndicBERTv2-MLM-Sam-TLM	0.376
IndicBERTv2-SS	0.338
IndicBERTv2-MLM-only	0.290

Table 1: Performance of IndicBERT-V2 model variants.

The IndicBERTv2-MLM-Back-TLM model performed the best, while the same base model without TF-IDF features scored a macro-F1 score of 0.362, implying the significance of TF-IDF features.

The class-wise precision, recall, and F1 score of our final model are presented in Table 2

The findings indicate that the IndicBERTv2-MLM-Back-TLM model was excellent in classifying the 'None' class which is easier to distinguish, with precision, recall, and F1 score of 1.00, 0.92, and 0.96, respectively. It did struggle with the 'negative' and 'substantiated' classes, which exhibit greater linguistic complexity.

Class	Precision	Recall	F1-score
Negative	0.13	0.15	0.14
Neutral	0.17	0.27	0.21
None	1.00	0.92	0.96
Opinionated	0.53	0.33	0.40
Positive	0.25	0.35	0.29
Sarcastic	0.50	0.42	0.45
Substantiated	0.22	0.24	0.23

Table 2: Class-wise performance in terms of Precision, Recall, and F1-score of our model.

Our final system, based on the IndicBERTv2-MLM-Back-TLM model, achieved a macro-F1 score of 0.3773, securing the 1st rank in the shared task. Table 3 shows the top 4 teams.

Project code files are available in Github.[1]

Rank	Team Name	Macro-F1
1	**Synapse**	**0.3773**
2	KCRL	0.3710
3	byteSizedLLM	0.3497
4	Eureka-CIOL	0.3187

Table 3: Top 4 teams in the shared task.

7 Conclusion

This paper studies political sentiment analysis on Tamil X (Twitter) comments using our custom model, which has achieved Rank 1 in this shared task. Our approach incorporates IndicBert-V2-MLM-Back-Translation and TF-IDF features, this addresses the challenges of informal, code-mixed discourse through pre-processing techniques such as demojizing and agent based hashtag expansion. Our approach can effectively capture linguistic nuances, making it suitably applicable to multi-class sentiment classification in political discussions. The study demonstrates the potential of NLP in analyzing public opinion in Tamil-speaking online communities and contributes to advancing sentiment analysis for Tamil language.

[1] https://github.com/SURIYA-KP/NAACL_DravidianLangTech_2025/

8 Limitations

Despite its effectiveness, our solution has certain limitations. The dataset used for training is relatively small, and its inherent ambiguity poses challenges, even for human interpretation. For instance, the query in tamil translating to "Naam Tamilar Vs AMMK.. Who will lead in Trichy?" is labelled as negative, despite its semantic proximity to a neutral, interrogative stance, thus limiting the model's ability to generalize across diverse political contexts. Highly ambiguous or sarcastic statements are also challenges for the model, in which the sentiment is hard to determine without more profound contextual understanding. Although IndicBERTv2 performs very well, even larger pretrained models with greater computational power could refine the accuracy of sentiment classification. Future research might address these aspects by using larger annotated datasets and further advanced transformer-based architectures.

References

A. Angdresey, L. Sitanayah, and I. L. H. Tangka. 2025. Sentiment analysis for political debates on youtube comments using bert labeling, random oversampling, and multinomial naïve bayes. *Journal of Computing Theories and Applications*, 2(3):342–354.

Prathvi B, Manavi K, Subrahmanyapoojary K, Asha Hegde, Kavya G, and Hosahalli Shashirekha. 2024. MUCS@DravidianLangTech-2024: A grid search approach to explore sentiment analysis in code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 257–261, St. Julian's, Malta. Association for Computational Linguistics.

B. R. Chakravarthi, R. Priyadharshini, S. Thavareesan, D. Chinnappa, D. Thenmozhi, E. Sherly, and C. Vasantharajan. 2021. Findings of the sentiment analysis of dravidian languages in code-mixed text. *arXiv preprint arXiv:2111.09811*.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the Shared Task on Political Multiclass Sentiment Analysis of Tamil X(Twitter) Comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. *Preprint*, arXiv:2212.05409.

S Sachin Kumar, M Anand Kumar, and KP Soman. 2017. Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets. In *Mining Intelligence and Knowledge Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings 5*, pages 320–334. Springer.

R. Rajalakshmi, S. Selvaraj, A. Shibani, and B. R. Chakravarthi. 2022. Understanding the role of emojis for emotion detection in tamil. In *Proceedings of the First Workshop on Multimodal Machine Learning in Low-resource Languages*, pages 9–17.

Prasanth S N, R Aswin Raj, Adhithan P, Premjith B, and Soman Kp. 2022. CEN-Tamil@DravidianLangTech-ACL2022: Abusive comment detection in Tamil using TF-IDF and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 70–74, Dublin, Ireland. Association for Computational Linguistics.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Abeera V P, Dr. Sachin Kumar, and Dr. Soman K P. 2023. Social media data analysis for Malayalam YouTube comments: Sentiment analysis and emotion detection using ML and DL models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

cuetRaptors@DravidianLangTech 2025: Transformer-Based Approaches for Detecting Abusive Tamil Text Targeting Women on Social Media

Md. Mubasshir Naib[a], Md. Saikat Hossain Shohag[b], Alamgir Hossain[c]
Jawad Hossain[d] and Mohammed Moshiul Hoque[e]
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904089[a], u1904088[b], u1704039[d]}@student.cuet.ac.bd
alamgir.hossain.cs@gmail.com[c], moshiul_240@cuet.ac.bd[e]

Abstract

With the exponential growth of social media usage, the prevalence of abusive language targeting women has become a pressing issue, particularly in low-resource languages (LRLs) like Tamil and Malayalam. This study is part of the shared task at DravidianLangTech@NAACL 2025, which focuses on detecting abusive comments in Tamil social media content. The provided dataset consists of binary-labeled comments (Abusive or Non-Abusive), gathered from YouTube, reflecting explicit abuse, implicit bias, stereotypes, and coded language. We developed and evaluated multiple models for this task, including traditional machine learning algorithms (Logistic Regression, Support Vector Machine, Random Forest Classifier, and Multinomial Naive Bayes), deep learning models (CNN, BiLSTM, and CNN+BiLSTM), and transformer-based architectures (DistilBERT, Multilingual BERT, XLM-RoBERTa), and fine-tuned variants of these models. Our best-performing model, Multilingual BERT, achieved a weighted F1-score of 0.7203, ranking 19th in the competition.

1 Introduction

The rapid expansion of social media has transformed communication, but it has also amplified the spread of abusive content, particularly targeting women and other marginalized groups (Priyadharshini et al., 2022b; Ghanghor et al., 2021b). In low-resource languages like Tamil, this issue is exacerbated by the lack of linguistic tools and datasets, making automated detection of abusive text a critical yet underexplored challenge (Chakravarthi et al., 2020; Priyadharshini et al., 2020). Tamil, a Dravidian language spoken by over 80 million people in South Asia, faces unique complexities due to its rich morphology, code-mixing tendencies, and the prevalence of implicit bias, stereotypes, and coded language in online discourse (Anita and Subalalitha, 2019; Subalalitha and Poovammal, 2018). While prior work has addressed abusive language detection in Tamil, most studies focus on broad categories (e.g., hate speech (Hossan et al., 2025), misogyny) or coarse-grained binary classification (Sharif et al., 2021b; Chakravarthi et al., 2022), with limited emphasis on nuanced abuse targeting women specifically.

Social media platforms like YouTube, Facebook, and Twitter have struggled to manually filter such content due to its sheer volume and linguistic diversity (Ghanghor et al., 2021a). Existing solutions for high-resource languages like English rely heavily on transformer-based models (Kumar et al., 2020; Sampath et al., 2022), but their efficacy in Tamil remains understudied. Recent initiatives like the DravidianLangTech shared tasks have spurred progress in abusive text detection (Chakravarthi et al., 2021; B et al., 2022), yet gaps persist in addressing gender-targeted abuse with computational efficiency and cultural sensitivity.

This work, part of the DravidianLangTech@NAACL 2025 shared task, focuses on detecting abusive Tamil social media comments directed at women. Besides using various ML and Dl models, We leverage transformer-based architectures—DistilBERT, Multilingual BERT (mBERT), and XLM-RoBERTa—to tackle binary classification on a dataset of YouTube comments labeled as *Abusive* or *Non-Abusive*. Our contributions include:

- A comparative analysis of traditional machine learning, deep learning, and lightweight transformer models for Tamil abuse detection.
- An evaluation of multilingual, language-specific pre-trained models and deep learning architectures (CNN, BiLSTM, CNN+BiLSTM) in capturing contextual and cultural nuances.

2 Related Task

The detection of abusive language in low-resource languages has gained traction in recent years, driven by the proliferation of harmful content on social media platforms.

Using classifiers like Logistic Regression, Support Vector Machines (SVM), and ensemble approaches, early attempts at abusive language detection concentrated on high-resource languages like English (Oswal, 2021). Traditional machine learning approaches have been the main method used in studies for low-resource languages like Bengali and Tamil. For example, (Eshan and Hasan, 2017) used SVM with tri-gram features to classify abusive Tamil texts with 95% accuracy. A weighted ensemble of BERT variants was also proposed by (Sharif and Hoque, 2021), who created a dataset of hostile Bengali text and achieved 93% weighted F1-scores. However, these studies rarely examine gender-specific abuse, instead concentrating on broad categories like hate speech and aggressiveness (Sharif et al., 2021a; Aurpa et al., 2021) or coarse-grained binary classification (e.g., abusive/non-abusive).

NLP jobs have been transformed by recent developments in transformer-based models, especially for high-resource languages. (Kumar et al., 2020), for instance, showed how effective BERT is in identifying implicit hate speech in English. However, morphological complexity, code-mixing, and cultural context make it difficult to adapt these models to low-resource languages like Tamil (Anita and Subalalitha, 2019). Although multilingual transformers (such as mBERT and XLM-RoBERTa) have demonstrated promise in cross-lingual tasks (Chakravarthi et al., 2021), nothing is known about how well they function in fine-grained abusive language detection, particularly when focused on women. Previous research in Devanagari script languages, including (Jha et al., 2020), used FastText to detect hate speech in Hindi with 92% accuracy, and (Chopra et al., 2023) used transformers to detect hate speech that was code-mixed between Hindi and English. These studies highlight the potential of hybrid and transformer-based approaches but underscore the need for language-specific adaptations.

Existing research on Tamil abusive language detection lacks focus on gender-targeted abuse and relies heavily on traditional ML methods (Priyadharshini et al., 2020; Chakravarthi et al., 2022). While (Sharif and Hoque, 2022) advanced Bengali aggression detection using BERT variants, similar efforts for Tamil are scarce. Our work bridges these gaps by:

Investigating transformer models (DistilBERT, mBERT, XLM-R) for detecting abusive Tamil text *targeting women*, a fine-grained and culturally sensitive task. Benchmarking against traditional ML baselines (Logistic Regression, Random Forest Classifier, and Multinomial Naive Bayes) and deep learning architectures (CNN, BiLSTM, CNN+BiLSTM) to quantify the benefits of lightweight transformers in low-resource settings. Addressing implicit bias and coded language through contextual embeddings, a challenge highlighted in prior Devanagari script research (Parihar et al., 2021; Nandi et al., 2024).

3 Task and Dataset Description

This shared task was organized to detect abusive Tamil and Malayalam texts targeting women on social media (Rajiakodi et al., 2025). The task focused on binary classification, categorizing texts as *Abusive* or *Non-Abusive*. We utilized the corpus provided by the organizers of Dravidian-LangTech@NAACL 2025 (Priyadharshini et al., 2023, 2022a), which comprises Tamil social media comments annotated for gender-specific abusive content. The dataset includes comments collected from YouTube, reflecting explicit abuse, implicit bias, stereotypes, and coded language targeting women. Table 1 summarizes the distribution of the dataset across training, validation, and test splits. While the dataset exhibits slight class imbalance, this reflects real-world social media data where abusive content often appears less frequently than non-abusive interactions.

Class	Train	Validation	Test	W_T	UW_T
Abusive	1236	129	305	25,585	13,181
Non-Abusive	1274	150	293	23,475	12,105
Total	2510	279	598	49,060	18,394

Table 1: Class distribution across training, validation, and test splits, where W_T represents total words and UW_T represents total unique words.

4 Dataset Visualization

Figure 1 represents the most common words in abusive texts in the training set, potentially indicating offensive or harmful language patterns. In contrast, Figure 2 highlights the frequent words

in non-abusive texts, reflecting more neutral and factual vocabulary. This analysis generated word clouds to visualize the linguistic characteristics of both categories, using a maximum of 200 words for each cloud, with word sizes proportional to their frequency.

Figure 1: Word Cloud distribution of *Abusive* class

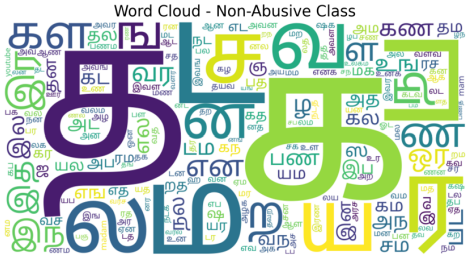


Figure 2: Word Cloud distribution of *Non-Abusive* class

5 Methodology

Various machine learning, deep learning, and transformer-based models were explored to establish baselines for detecting abusive language in Tamil comments in Figure 3. The implementation details of the models have been open-sourced to ensure reproducibility[1] .

5.1 Data Preprocessing

The dataset was preprocessed to ensure quality and consistency by removing rows with missing or invalid values in the "Class" column, mapping binary labels ("Abusive" to 1, "Non-Abusive" to 0), and verifying the absence of NaN values. It was then split into training and validation sets using a stratified split to maintain label distribution, resulting in a clean and balanced dataset ready for modeling.

5.2 Feature Extraction

To represent text data numerically, TF-IDF, Bag of Words (BoW), and FastText embeddings were used.

[1]https://github.com/MubasshirNaib/Detecting-Abusive-Tamil-Text

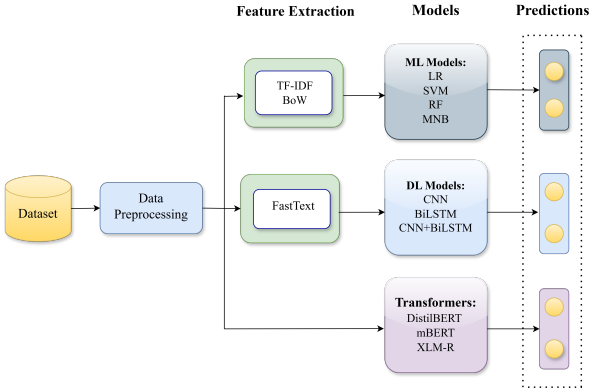


Figure 3: Schematic process for detecting abusive comments in Tamil social media content.

TF-IDF and BoW extracted the top 5,000 features to capture word importance and occurrences. FastText embeddings, trained on the tokenized training data with 100-dimensional vectors, provided context-aware representations, enhancing the ability of models to capture linguistic nuances. These methods ensured diverse and effective feature representations for model evaluation.

5.3 Machine Learning Models

For this task, we evaluated multiple machine learning models to detect abusive comments, including Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier, and Multinomial Naive Bayes (MNB). By training and evaluating these models, we compared their capabilities comprehensively, enabling a deeper understanding of their strengths and suitability for the classification task.

5.4 Deep Learning Models

For this task, we implemented three deep learning models—CNN, BiLSTM, and a hybrid CNN+BiLSTM—for this classification. The CNN model includes a 1D convolution layer with 128 filters and ReLU activation, followed by max-pooling and fully connected layers. The BiLSTM model uses a bidirectional LSTM layer with 128 units to capture long-term dependencies from both directions. The combined CNN+BiLSTM model integrates two convolutional layers followed by a max-pooling and BiLSTM layer, leveraging both local feature extraction and sequential context learning. A dropout rate of 0.5 is applied to reduce overfitting, and models are trained using the Adam optimizer with binary cross-entropy loss for five epochs, ensuring robust prediction performance.

5.5 Transformer-Based Models

Transformer-Based models are particularly well-suited for multilingual and cross-lingual tasks, making them ideal for addressing abusive language detection in low-resource languages like Tamil. To tackle the shared task, we experimented with various transformer-based architectures, including DistilBERT (Sanh et al., 2020), Multilingual BERT (m-BERT) (Pires et al., 2019), and XLM-RoBERTa (XLM-R) (Conneau et al., 2020). Each model was fine-tuned on the binary classification task of identifying abusive and non-abusive comments in Tamil social media data. Here, the multilingual BERT was fine-tuned using the following hyperparameters shown in the Table 2. These hyperparameter choices ensured a balance between convergence and regularization, enabling the model to achieve a weighted F1-score of 0.7203. This demonstrates Multilingual BERT's ability to effectively capture nuanced patterns of abusive language while maintaining computational efficiency.

Parameter	Value
Batch Size	16
Epochs	7
Weight Decay	0.003
Learning Rate	5e-5

Table 2: Hyperparameters used in the best model

6 Results and Analysis

The performance of the various methods is presented in Table 3 The macro F1-score is used to evaluate and compare the overall performance of the models. Among the traditional machine learning models, Logistic Regression (LR) achieved the highest performance with an F1-score of 0.6933, an accuracy of 0.6935, and a G1-Score of 0.6833, outperforming both SVM and RF. The SVM model, while showing competitive results, lagged behind LR with an F1-score of 0.6746, an accuracy of 0.6756, and a G1-Score of 0.6764. Random Forest (RF) showed consistent performance but did not surpass LR or SVM, achieving an F1-score of 0.6738, an accuracy of 0.6738, and a G1-Score of 0.6739.

Deep learning models such as CNN and CNN+BiLSTM showed moderate performance, with an F1-score of 0.5679 for CNN and 0.5680 for CNN+BiLSTM, both having a G1-Score of 0.5681. BiLSTM, on the other hand, had a significantly lower performance with an F1-score of 0.3294, an accuracy of 0.4964, and a G1-Score of 0.3497.

Among the transformer models, m-BERT achieved the highest F1-score of 0.7203, an accuracy of 0.6404, and a G1-Score of 0.7233, followed by DistilBERT with an F1-score of 0.7068, an accuracy of 0.6164, and a G1-Score of 0.7183. XLM-R demonstrated a strong recall of 1.0000 but delivered lower overall performance with an F1-score of 0.6521, an accuracy of 0.4838, and a G1-Score of 0.6656.

Classifier	P	R	F1	A	G1
LR	0.68	0.68	0.68	0.68	0.68
SVM	0.67	0.67	0.67	0.67	0.67
RF	0.67	0.67	0.67	0.67	0.67
MNB	0.69	0.69	0.69	0.69	0.69
CNN	0.56	0.56	0.56	0.56	0.56
BiLSTM	0.24	0.49	0.32	0.49	0.35
CNN+BiLSTM	0.56	0.56	0.56	0.56	0.56
m-BERT	**0.59**	**0.96**	**0.72**	**0.64**	**0.72**
DistilBERT	0.56	0.93	0.70	0.61	0.71
XLM-R	0.48	1.00	0.65	0.48	0.66

Table 3: Performance of various models, where P, R, F1, A and G1 denote precision, recall, macro F1-score, accuracy and G1-Score respectively.

Overall, transformer-based models, particularly Multilingual BERT (m-BERT), excelled due to its pretraining on a multilingual corpus, including Tamil, enabling it to grasp contextual nuances of abusive language. Its self-attention mechanism outperforms traditional models (e.g., Logistic Regression, SVM), which miss subtleties, and deep learning models (e.g., CNN, BiLSTM), which struggle with limited data or long-range dependencies. While m-BERT uses generalized embeddings rather than Tamil-specific ones, its Tamil exposure was enough for strong performance (F1: 0.7203). Tamil-specific embeddings might enhance results but this is not explored in this work.

6.1 Error Analysis

We conducted both quantitative and qualitative error analyses to gain comprehensive insights into the performance of the proposed model.

6.1.1 Quantitative Analysis:

The classifier demonstrated notable performance in identifying abusive and non-abusive content. However, a closer examination of the confusion matrix, Figure 4 reveals key areas of error, providing in-

sights into the model's behavior across the different classes.

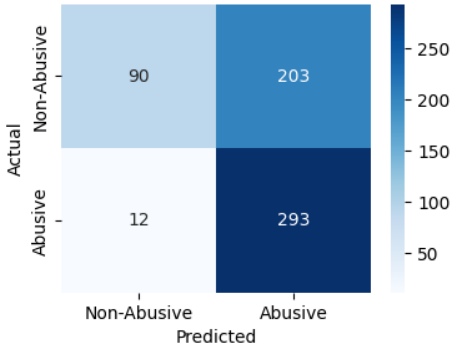


Figure 4: Confusion matrix of m-Bert

The classifier performed well in identifying abusive content, achieving a high True Positive Rate (TPR) of 96.05% for the abusive class, with minimal misclassification. However, the non-abusive class had a significantly lower TPR of 30.72%, with a large number of non-abusive instances being incorrectly classified as abusive. This suggests an overprediction of the abusive class, potentially caused by class imbalance, ambiguous features, or limited representation of non-abusive examples in the training data. To improve performance, the issues of class imbalance and feature ambiguity need to be addressed by refining the dataset, enhancing feature representation, and employing better modeling techniques to improve the classification of non-abusive content while maintaining high recall for the abusive class.

6.1.2 Qualitative Analysis:

Figure 5 illustrates a qualitative analysis of the m-BERT model's predictions for the abusive language detection task. The model correctly classified samples 1 and 5 as *Abusive* and samples 3 and 4 as *Non-Abusive*, demonstrating its effectiveness in distinguishing between different language tones. However, sample 2 was misclassified as *Abusive* instead of *Non-Abusive*, likely due to contextual ambiguity or overlapping linguistic patterns in the dataset. This misclassification highlights a potential area for improvement in capturing subtle differences in expression.

7 Conclusion

This study explored a range of machine learning, deep learning, and transformer-based models for detecting abusive language in Tamil social media content. Among these, m-BERT emerged as the best-performing model, achieving a macro F1-score of 0.7203, showcasing its effectiveness in capturing nuanced patterns in text. Transformer-based models demonstrated clear advantages over traditional and deep learning approaches, highlighting their ability to manage complex tasks like abusive language detection. This study underscores the importance of leveraging advanced models and fine-tuning strategies to improve the detection of abusive content in low-resource, code-mixed languages.

Sample Text	Actual Label	Predicted Label
Sample Text 1: இவ ஒரு மானெங்கெட்ட பொறுக்கி. ஒரே ஒரு routine ஒர்க் அவளுக்கு இருக்குறது தண்ணிய போட்டுட்டு அசிங்கமா பேசுறது.	**Abusive**	**Abusive**
Sample Text 2: இப்டியே பேசிக்கிட்டே இருந்தா எப்டி..... யாரு பெருசுனு அடிசிக்காட்டு ...	**Non- Abusive**	**Abusive**
Sample Text 3: அடக் கடவுளே இது என்னக் கொடுமையை ஊருல உலகத்துல எவ்வளவு பிரச்சினை இருக்கு இது என்னக்கொடுமை அடேய் கார்த்திக் நீ எங்கு இருந்தாலும் வந்துவிடு உன் காளில் விழுகின்றேன்	**Non- Abusive**	**Non- Abusive**
Sample Text 4: இதற்கு ஒரு தீர்வு இருக்கு. அவன் அவன் வேலை அவன் அவன் பார்த்தால் எந்த பாதிப்பும் ஏற்படாது.	**Non- Abusive**	**Non- Abusive**
Sample Text 5: தம்பி போய் நல்லவங்களை பேட்டிஎடு அவ சொல்வது அத்தனையும் பொய் தெரியாதா உனக்கு	**Abusive**	**Abusive**

Figure 5: Some outputs predicted by the best model(m-Bert).

Limitations

Despite the success of m-BERT, the system exhibited an overprediction tendency for the abusive class, struggling to accurately classify non-abusive content. This imbalance reflects challenges related to skewed class distribution, feature ambiguity, and limited representation of non-abusive data in the training set. Additionally, the reliance on pre-trained transformer models restricted opportunities for domain-specific optimization. Addressing these limitations will require balancing datasets, employing data augmentation strategies, and exploring innovative model architectures tailored to the complexities of low-resource, code-mixed languages like Tamil.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

R. Anita and C.N. Subalalitha. 2019. An approach to cluster tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4.

Tanjim Taharat Aurpa, Rifat Sadik, and Md Shoaib Ahmed. 2021. Abusive bangla comments detection on facebook using transformer-based deep learning models. *Social Network Analysis and Mining*, 12(1):24.

Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya N, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 339–345, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Philip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *CoRR*, abs/2109.00227.

Abhishek Chopra, Deepak Kumar Sharma, Aashna Jha, and Uttam Ghosh. 2023. A framework for online hate speech detection on code-mixed hindi-english text and hindi text in devanagari. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(5).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Shahnoor C. Eshan and Mohammad S. Hasan. 2017. An application of machine learning to detect abusive bengali text. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021a. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

Nikhil Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021b. IIITK@LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil , Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 197–203, Kyiv. Association for Computational Linguistics.

Md. Refaj Hossan, Nazmus Sakib, Md. Alam Miah, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET_Big_O@NLU of Devanagari script languages 2025: Identifying script language and detecting hate speech using deep learning and transformer model. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 253–259, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Vikas Kumar Jha, Hrudya P, Vinu P N, Vishnu Vijayan, and Prabaharan P. 2020. Dhot-repository and classification of offensive tweets in the hindi language. *Procedia Computer Science*, 171:2324–2333. Third International Conference on Computing and Network Communications (CoCoNet'19).

Ritesh Kumar, Atul Kr. Ojha, Bornini Lahiri, Marcos Zampieri, Shervin Malmasi, Vanessa Murdock, and Daniel Kadar, editors. 2020. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA), Marseille, France.

Arpan Nandi, Kamal Sarkar, Arjun Mallick, and Arkadeep De. 2024. A survey of hate speech detection in indian languages. *Social Network Analysis and Mining*, 14(1):70.

Nikhil Oswal. 2021. Identifying and categorizing offensive language in social media. *CoRR*, abs/2104.04871.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022a. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2022b. Overview of the dravidiancodemix 2021 shared task on sentiment detection in tamil, malayalam, and kannada. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 4–6, New York, NY, USA. Association for Computing Machinery.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P. McCrae. 2020. Named entity recognition for code-mixed indian corpus using meta embedding. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 68–72.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Cn, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, Kishore Ponnusamy, and Santhiya Pandiyan. 2022. Findings of the shared task on emotion analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 279–285, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Omar Sharif and Mohammed Moshiul Hoque. 2021. Identification and classification of textual aggression in social media: Resource creation and evaluation. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 9–20, Cham. Springer International Publishing.

Omar Sharif and Mohammed Moshiul Hoque. 2022. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*, 490:462–481.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021a. Combating hostility: Covid-19 fake news and hostile post detection in social media. *CoRR*, abs/2101.03291.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021b. NLP-CUET@DravidianLangTech-EACL2021: Offensive language detection from multilingual code-mixed text using transformers. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 255–261, Kyiv. Association for Computational Linguistics.

C.N Subalalitha and E. Poovammal. 2018. Automatic bilingual dictionary construction for tirukural. *Applied Artificial Intelligence*, 32(6):558–567.

KEC_AI_BRIGHTRED@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages

Kogilavani Shanmugavadivel[1] **, Malliga Subramanian**[2]**,**
Nishdharani P[1]**,Santhiya E**[1]**,Yaswanth Raj E**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{nishdharanip05, santhiyae587, yaswanthraje2004}@gmail.com

Abstract

Hate speech detection in multilingual settings presents significant challenges due to linguistic variations and speech patterns across different languages. This study proposes a fusion-based approach that integrates audio and text features to enhance classification accuracy in Tamil, Telugu, and Malayalam. We extract Mel-Frequency Cepstral Coefficients and their delta variations for speech representation, while text-based features contribute additional linguistic insights. Several models were evaluated, including BiLSTM, Capsule Networks with Attention, Capsule-GRU, ConvLSTM-BiLSTM, and Multinomial Naïve Bayes, to determine the most effective architecture. Experimental results demonstrate that Random Forest performs best for text classification, while CNN achieves the highest accuracy for audio classification. The model was evaluated using the Macro F1 score and ranked ninth in Tamil with a score of 0.3018, ninth in Telugu with a score of 0.251, and thirteenth in Malayalam with a score of 0.2782 in the Multimodal Social Media Data Analysis in Dravidian Languages shared task at DravidianLangTech@NAACL 2025. By leveraging feature fusion and optimized model selection, this approach provides a scalable and effective framework for multilingual hate speech detection, contributing to improved content moderation on social media platforms.

1 Introduction

With the rise of social media and digital communication, hate speech has become a major concern, particularly in multilingual communities. Traditional hate speech detection methods primarily rely on text analysis, but spoken content, such as audio messages and voice notes, also plays a crucial role in spreading harmful discourse. Detecting hate speech in languages like Tamil, Telugu, and Malayalam presents unique challenges due to code-mixing, informal language structures, and phonetic variations.

This study addresses these challenges by incorporating both text and audio-based features to improve classification accuracy. We extract MFCC and delta features from speech data and apply various deep learning and machine learning models to analyze textual content. By evaluating models such as BiLSTM, Capsule Networks, GRU, ConvLSTM, and Naïve Bayes, we identify the most effective approach for each modality. Our results demonstrate that Random Forest performs best for text-based hate speech detection, while CNN excels in audio-based classification. This research contributes to enhancing multilingual content moderation by leveraging both acoustic and linguistic features for more robust hate speech detection.

2 Literature Survey

(Lal G et al., 2025), presented an overview of the shared task on multimodal hate speech detection in Tamil, Telugu, and Malayalam at DravidianLangTech@NAACL 2025. It discusses dataset creation, preprocessing techniques, and model performance in identifying hate speech across text and audio modalities.(Premjith et al., 2024) analyzed multimodal social media data, including text, audio, and video from platforms like Twitter and YouTube. Their study focused on sentiment analysis, abusive language detection, and hate speech detection. The results were presented for Dravidian languages.(Sreelakshmi et al., 2024) showed that MuRIL embeddings with an SVM (RBF kernel) performed well across six datasets. The highest accuracies were 66% (Kannada), 72% (Tamil), and 96% (Malayalam) for DravidianLangTech 2021. HASOC 2021 achieved 68% (Malayalam) and 76% (Tamil), while HASOC 2020 reached 92% (Malayalam). (Mohan et al., 2023) proposed a multimodal approach for hate speech detection in Tamil us-

ing TimeSformer for video, Wav2vec2 for audio, and BERT-based models for text. They achieved 81.82% accuracy (F1: 68.65%) for text, 63.63% accuracy (F1: 50.60%) for audio, and 45.45% accuracy (F1: 33.64%) for video. By combining features from all modalities, they achieved 81.82% accuracy and a 66.67% F1 score. (Arunachalam and Maheswari, 2024) proposed a method to detect hateful remarks in Dravidian languages on social media. Using mBERT with CATBOOST and GSCV, they achieved F1 scores of 0.94 (Tamil), 0.98 (Malayalam), and 0.82 (Kannada) on the Dravidian Code-Mix FIRE 2021 dataset. Their approach effectively analyzed YouTube comments using various preprocessing techniques and binary classifiers. (Roy et al., 2022) developed a deep ensemble framework using deep learning and transformer models to detect offensive posts in Tamil-Malayalam code-mixed text. Their approach achieved weighted F1-scores of 0.802 (Malayalam) and 0.933 (Tamil). The model outperformed state-of-the-art methods on these datasets. (Dhanya and Balakrishnan, 2021) promoted the creation of an automated hate speech detection system for Malayalam by presenting a survey.

3 Task Description

The task aims to develop an effective multimodal hate speech detection system that can process and classify hate speech in both textual and audio formats across Tamil, Telugu, and Malayalam. The challenge is divided into three key components: the first focuses on detecting hate speech from textual data by classifying transcripts into "hate speech" or "non-hate speech." The second component deals with audio-based classification, where audio features (e.g., MFCCs, spectral features) are extracted and used to identify hate speech. The core of the task involves multimodal fusion, where the outputs of text-based classification using Random Forest and audio-based classification using CNN are combined to enhance overall detection accuracy. The models will be evaluated using metrics such as accuracy, precision, recall, and F1-score on both individual and multimodal data.

4 Dataset description

The dataset used for this task consists of multimodal data from social media platforms, specifically targeting hate speech detection in Tamil, Telugu, and Malayalam. It contains both text and audio data, with each instance representing a piece of social media content that could potentially contain hate speech.

Language	Non-Hate	Hate (C,G,P,R)
Malayalam	406	477
Tamil	287	227
Telugu	198	358

Table 1: Text Data Distribution

Language	Non-Hate	Hate (C,G,P,R)
Malayalam	406	477
Tamil	287	222
Telugu	198	353

Table 2: Audio Data Distribution

The dataset is categorized into two main classes: Hate and Non-Hate. The hate speech instances are further classified into four subclasses based on their nature: Gender (G), Political (P), Religious (R), and Personal Defamation (C).

5 Methodology

In this section, we outline the approach taken for multimodal hate speech detection using both text and audio data.

5.1 Preprocessing Data

5.1.1 Text

Preprocessing Tamil, Malayalam, and Telugu texts involves data cleaning (removal of special characters, numbers, and extra spaces), tokenization using language-specific tools, normalization for spelling variations, and case conversion. TF-IDF represents text numerically, while stemming or lemmatization reduces words to root forms. SMOTE addresses class imbalance by generating synthetic samples. These steps ensure a clean and balanced dataset for effective hate speech detection.

5.1.2 Audio

Preprocessing Tamil, Telugu, and Malayalam audio involves normalization for consistent volume, noise reduction using spectral gating, and resampling (e.g., 16 kHz). Silence trimming removes pauses, and phoneme segmentation improves accuracy. Key acoustic features like MFCCs and Mel Spectrograms are extracted, while speaker normalization minimizes variability. Data augmentation

(e.g., pitch shifting, time-stretching) enhances robustness, ensuring effective hate speech detection in Dravidian languages.

5.2 Models Developed and Evaluated

We explored and compared several models to address the task of multimodal hate speech detection.

5.2.1 Random Forest (Text Classification)

We implemented a Random Forest classifier, an ensemble method known for handling noisy data and capturing complex feature relationships. It achieved the highest accuracy on text data. The model was trained on extracted text features, demonstrating strong performance.

5.2.2 Convolutional Neural Network (CNN) (Audio Classification)

For audio classification, a CNN was used to process MFCC-extracted features, capturing spatial hierarchies and detecting hate speech patterns. The model showed high accuracy. This section outlines the approach for multimodal hate speech detection using text and audio data.

5.2.3 Bi-directional LSTM (BiLSTM)

We experimented with BiLSTM networks for text but found they underperformed compared to the Random Forest model. While BiLSTM captures long-range dependencies, it did not generalize well for hate speech detection in this dataset.

5.2.4 Capsule Networks with Attention-based BiLSTM

We evaluated a Capsule Network with Attention-based BiLSTM to capture spatial hierarchies and key features. Despite its theoretical benefits, it did not outperform the simpler Random Forest or CNN models.

5.2.5 Capsule Networks with GRU

We tested a Capsule Network with GRU, leveraging GRUs for sequential data processing. However, the integration did not improve accuracy, and the model performed worse than others.

5.2.6 ConvLSTM + BiLSTM

We tested the ConvLSTM + BiLSTM model, combining Convolutional LSTM with BiLSTM to capture spatial and temporal dependencies. However, its complexity led to overfitting on the training data. As a result, it performed worse than simpler models with lower accuracy.

5.2.7 Multinomial Naive Bayes (Text Classification)

The Multinomial Naive Bayes model for text classification, it performed poorly due to its inability to handle data noise and its assumption of feature independence, making it unsuitable for multimodal hate speech detection.

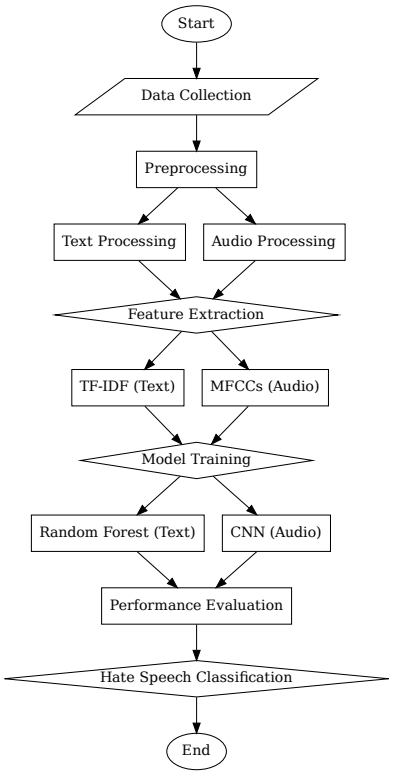


Figure 1: Proposed Model Workflow

5.3 Model Selection

The Random Forest model for text and the CNN model for audio were chosen because of their higher classification task accuracy. Convolutional Neural Networks (CNNs) greatly improved the detection of hate speech in spoken content by exhibiting exceptional efficiency in extracting crucial audio information, such as temporal fluctuations and frequency patterns. CNNs were able to capture complicated audio representations that were frequently missed by standard models by utilizing many layers of feature extraction.

On the other hand, because of its capacity to handle high-dimensional data and capture linguistic subtleties, the Random Forest model demonstrated remarkable efficacy in text classification. By combining several decision trees, Random For-

est's ensemble learning feature enabled strong generalization and enhanced resistance to overfitting. This made it particularly adept at distinguishing subtle variations in textual content, such as sarcasm, implicit biases, and contextual dependencies—factors that are crucial for accurately identifying hate speech in written form.

5.4 Performance Comparison

5.4.1 Text Classification:

The Random Forest classifier performed the best in terms of training accuracy, surpassing BiLSTM and Naive Bayes.

Class/Metric	Precision	Recall	F1-Score
Tamil (Text)			
C	0.93	0.93	0.93
G	0.88	0.89	0.89
N	1.00	0.98	0.99
P	1.00	0.98	0.99
R	0.98	0.95	0.96
Accuracy	-	-	0.94
Macro Avg	0.94	0.94	0.94
Weighted Avg	0.94	0.94	0.94
Telugu (Text)			
C	0.89	0.95	0.92
G	0.74	0.97	0.84
N	0.85	0.79	0.82
P	1.00	0.88	0.93
R	0.97	0.84	0.90
Accuracy	-	-	0.89
Macro Avg	0.89	0.89	0.89
Weighted Avg	0.90	0.88	0.89
Malayalam (Text)			
C	0.87	0.87	0.87
G	0.94	0.88	0.91
N	0.55	0.79	0.65
P	1.00	0.73	0.85
R	1.00	0.73	0.85
Accuracy	-	-	0.82
Macro Avg	0.87	0.82	0.83
Weighted Avg	0.87	0.82	0.83

Table 3: Detailed Classification Report for Tamil, Telugu, and Malayalam (Text)

5.4.2 Audio Classification:

The CNN model outperformed all other deep learning models, including BiLSTM and ConvLSTM-based models, which struggled with audio data.

Class/Metric	Precision	Recall	F1-Score
Tamil (Audio)			
C	1.00	0.20	0.33
G	0.00	0.22	0.36
P	1.00	0.58	0.73
R	0.77	0.77	0.77
N	0.67	0.96	0.79
Accuracy	-	-	0.71
Macro Avg	0.85	0.53	0.58
Weighted Avg	0.76	0.71	0.66
Telugu (Audio)			
C	0.70	0.64	0.67
G	0.71	0.50	0.59
N	0.68	0.90	0.77
P	0.91	0.67	0.77
R	0.80	0.57	0.67
Accuracy	-	-	0.72
Macro Avg	0.76	0.69	0.72
Weighted Avg	0.73	0.72	0.71
Malayalam (Audio)			
C	0.00	0.00	0.00
G	0.33	0.11	0.17
P	0.33	0.17	0.22
R	0.78	0.54	0.64
N	0.63	0.96	0.76
Accuracy	-	-	0.62
Macro Avg	0.42	0.36	0.36
Weighted Avg	0.52	0.62	0.53

Table 4: Detailed Classification Report for Tamil, Telugu, and Malayalam (Audio)

5.5 Performance Evaluation

The performance of the various models used for multimodal hate speech detection, comparing the accuracy of both text and audio classification models. The models explored include BiLSTM, Capsule-based models, ConvLSTM, Multinomial Naive Bayes, Random Forest and CNN. This section presents accuracy results for text models in Tamil, Telugu, and Malayalam. Random Forest outperformed all models, achieving the highest accuracy in Tamil (0.9373), Telugu (0.8838), and Malayalam (0.8153). Multinomial Naive Bayes performed reasonably but was outperformed by Random Forest. BiLSTM and other models like Capsule + Attention-based BiLSTM and ConvLSTM + BiLSTM showed poor performance, with BiLSTM scoring just 0.087 across all languages. CNN for Audio achieved good accuracy, with the highest in Malayalam (0.7577), followed by Telugu (0.7207) and Tamil (0.7059).

Model	Tamil	Telugu	Malayalam
Random Forest (Text)	0.9373	0.8838	0.8153
CNN (Audio)	0.7059	0.7207	0.7577
Capsule+Attention based BiLSTM (Text)	0.5922	0.5434	0.5263
Capsule+GRU (Text)	0.5437	0.5260	0.5132
ConvLSTM+ BiLSTM(Text)	0.5340	0.5421	0.6051
Multinomial Naive Bayes (Text)	0.6699	0.7021	0.6901

Table 5: Accuracy Comparison

6 Limitations

There are limitations to this study that need more research. Performance could be improved by enhancing the fusion strategy with transformer-based solutions. Optimization is required, according to the F1-scores (Tamil: 0.3018, Telugu: 0.251, Malayalam: 0.2782). Complex designs such as ConvLSTM and BiLSTM performed worse than simpler models. Future research should improve models, hone fusion techniques, compare to the most advanced methods, and guarantee reproducibility through open-source implementation.

7 Conclusion

In conclusion, the hate speech recognition system for Tamil, Telugu, and Malayalam showed promising results by integrating preprocessing techniques for both audio and text. The Random Forest model effectively captured semantic features for text, achieving high accuracy, while the CNN model extracted key features from audio signals, also yielding high accuracy. The fusion of these models enhanced performance, enabling more accurate and context-aware predictions by utilizing both modalities. The source code for our approach is available at https://github.com/NishdharaniP/Multimodal_hatespeech_detection.git.

References

V Arunachalam and N Maheswari. 2024. Enhanced detection of hate speech in dravidian languages in social media using ensemble transformers. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19:036.

LK Dhanya and Kannan Balakrishnan. 2021. Hate speech detection in asian languages: a survey. In *2021 international conference on communication, control and information sciences (ICCISc)*, volume 1, pages 1–5. IEEE.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Jayanth Mohan, Spandana Reddy Mekapati, and Bharathi Raja Chakravarthi. 2023. A multimodal approach for hate and offensive content detection in tamil: From corpus creation to model development. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

Author Index