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

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 121–125
May 3, 2025 ©2025 Association for Computational Linguistics

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

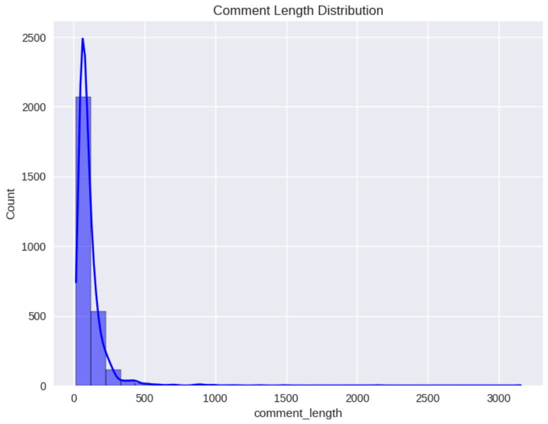

Figure 1: Distribution of comment lengths in Tamil language dataset.

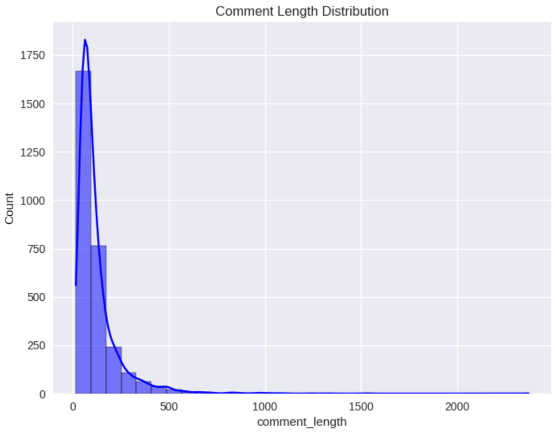

Figure 2: Distribution of comment lengths in Malayalam language dataset.

Word Cloud for All Comments
இந்த
என்ன
தான்
நீ
போய்
மாதிரி
எல்லாம்
அந்த
கார்த்தி
மேடம்
இருக்கு
இல்ல
உங்க
நீங்க
இவ
நல்ல
என்று
கார்த்திக்
அம்மா
கூட
வந்து

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 DravidianLangTech 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-

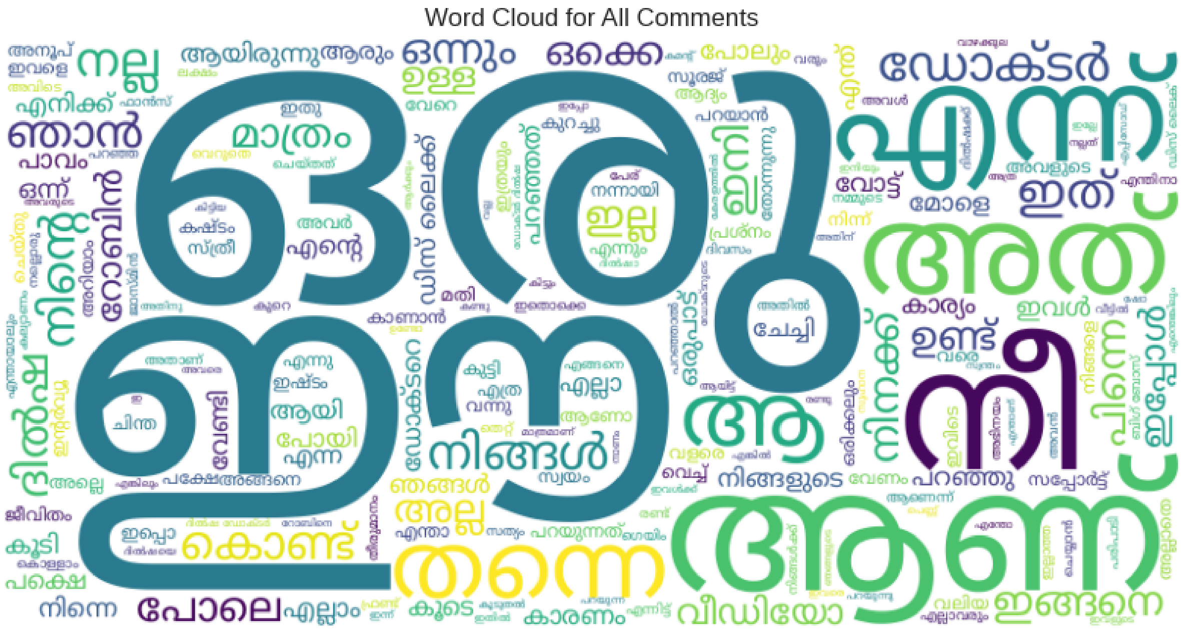

Figure 4: Word cloud representing all comments in Malayalam.

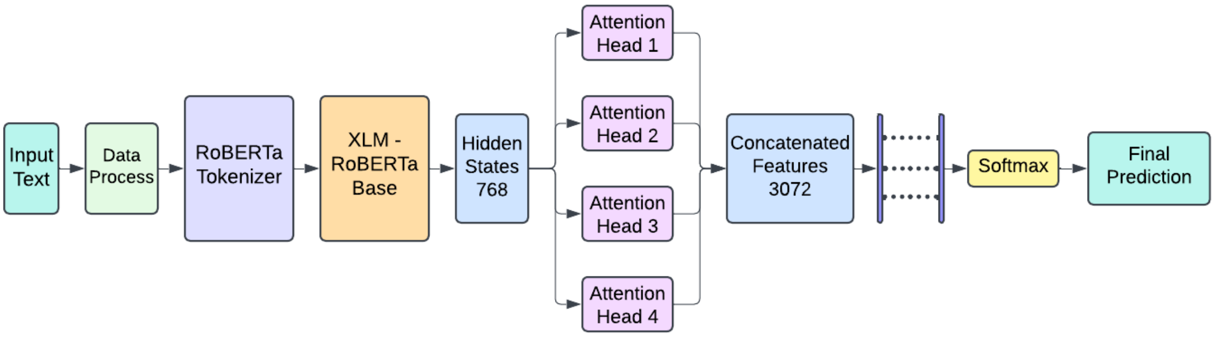

Figure 5: Proposed Model Architecture

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.