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 **3$^{\text{rd}}$** 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-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 561–566
May 3, 2025 ©2025 Association for Computational Linguistics

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.

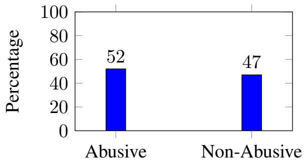


Figure 1: Statistics of training dataset.

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.

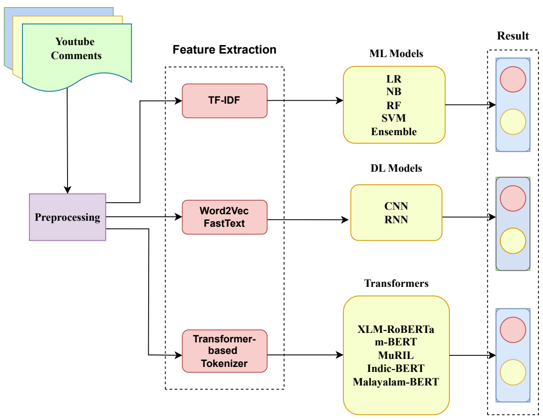


Figure 2: Abstract view of our methodology.

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 hyperparameters, 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.

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.

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 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.

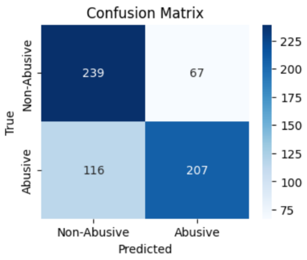


Figure 3: Confusion matrix of our best performing model.

5.2 Qualitative Discussion

Table 4 highlights both correctly classified and misclassified cases. Among the misclassified cases:

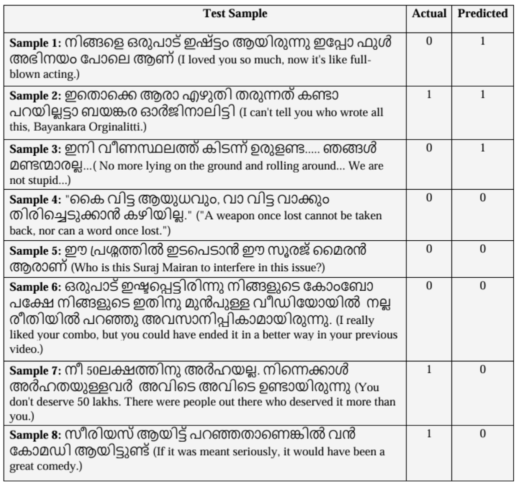
False Positives:

- **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.

These misclassifications indicate the model's struggle with indirect abuse and sarcasm.



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.

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.