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 details 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.

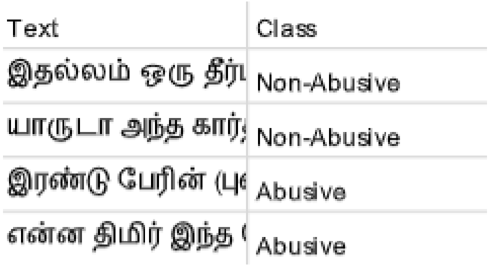
In recent work, transformer-based models have been put into the centre of attention instead of conventional models, and they have been shown to yield better performance in text classification tasks. (Vaswani et al., 2017) described the Transformer architecture, which has served as the basis for various 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.

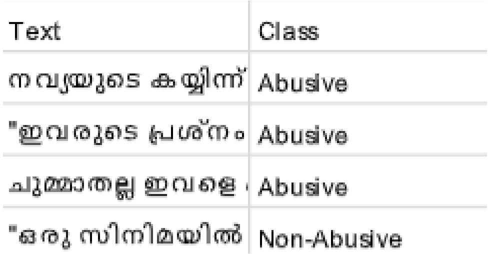
3 Task Description

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 தீர	Non-Abusive
யாருடா அந்த கார்	Non-Abusive
இரண்டு பேரின் (பு	Abusive
என்ன திமிர் இந்த	Abusive

Figure 1: Tamil Abusive language Dataset



Text	Class
നവ്യയുടെ കയ്യിന്	Abusive
"ഇവരുടെ പ്രശ്നം	Abusive
ചുമ്മാതല്ല ഇവളെ	Abusive
"ഒരു സിനിമയിൽ	Non-Abusive

Figure 2: Malayalam Abusive language Dataset

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 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.

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 pretraining 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.