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 harassments 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-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 493–497
May 3, 2025 ©2025 Association for Computational Linguistics

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.

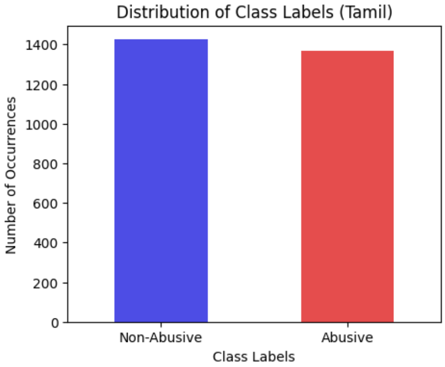

Figure 1: Dataset Description - Distribution of abusive and non-abusive comments in Tamil

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 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.

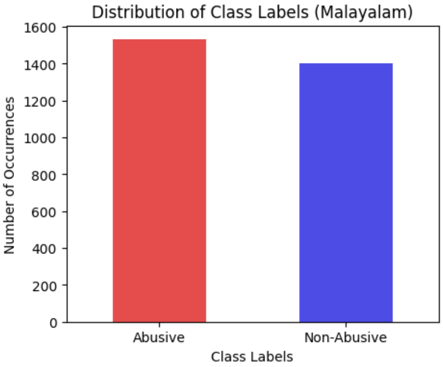

Figure 2: Dataset Description - Distribution of abusive and non-abusive comments in Malayalam

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 learning 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.