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 pretrained 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

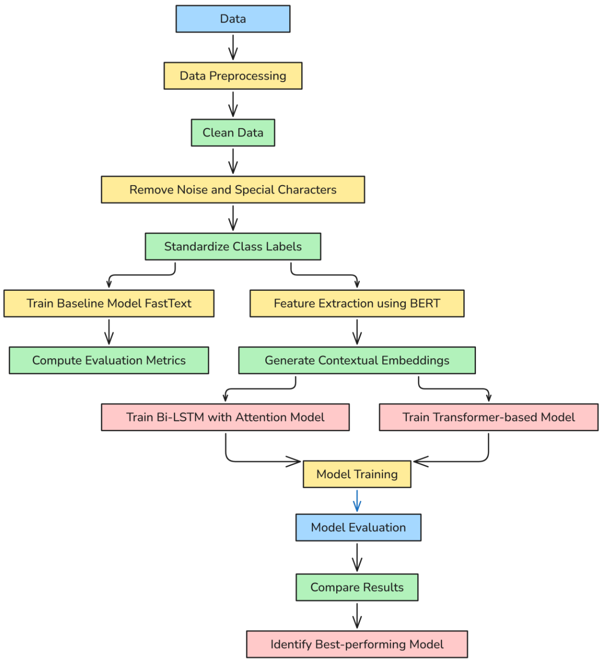
Data
Data Preprocessing
Clean Data
Remove Noise and Special Characters
Standardize Class Labels
Train Baseline Model FastText
Feature Extraction using BERT
Compute Evaluation Metrics
Generate Contextual Embeddings
Train Bi-LSTM with Attention Model
Train Transformer-based Model
Model Training
Model Evaluation
Compare Results
Identify Best-performing Model

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

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

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 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.