SSN_IT_NLP@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media

Maria Nancy C[1], **Radha N** [2], **Swathika R**[3]
[1]Annai Veilankanni's College of Engineering, Nedungundram , India
[2,3] Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, India
nancycse13@gmail.com[1]
radhan@ssn.edu.in[2]
swathikar@ssn.edu.in[3]

Abstract

The proliferation of social media platforms has led to a rise in online abuse, particularly against marginalized groups such as women. This study focuses on the classification of abusive comments in Tamil and Malayalam, two Dravidian languages widely spoken in South India. Leveraging a multilingual BERT model, this paper provides an effective approach for detecting and categorizing abusive and non-abusive text. Using labeled datasets comprising social media comments, our model demonstrates its ability to identify targeted abuse with promising accuracy. This paper outlines the dataset preparation, model architecture, training methodology, and the evaluation of results, providing a foundation for combating online abuse in low-resource languages. This methodology uniquely integrates multilingual BERT and weighted loss functions to address class imbalance, paving the way for effective abuse detection in other underrepresented languages. The BERT model achieved an F1-score of 0.6519 for Tamil and 0.6601 for Malayalam. The code for this work is available on github Abusive-Text-targeting-women.

1 Introduction

Social media platforms have grown to be an important online forum for entertainment, communication, and information exchange in recent years. Despite their advantages, these platforms are increasingly misused to target women with derogatory language. Due to cultural biases and gender inequality, women are frequently the victim of cruel and disparaging remarks that aim to denigrate, harass, or threaten them. Women may face significant psychological, social, and professional repercussions from this form of online abuse, a distinct type of cyberbullying that necessitates appropriate intervention. By identifying offensive language directed at women in comments, we address this issue by concentrating on online content management. In order to complete this assignment, we used targeted searches to scrape YouTube comments on sensitive and contentious subjects where gender-based abuse is common(Rajiakodi et al., 2025).These queries targeted explicit abuse, implicit bias, stereotypes, and coded language. This task's objective is to determine whether or not a certain comment contains abusive language. Text in the low-resource South Indian languages of Tamil and Malayalam is included in the dataset.

2 Related Work

The fast growth of social media has amplified the existence of online abuse, which targets more women as they mainly represent the underprivileged group and suffer more and in different ways (Chakravarthi et al., 2023). However, the two Dravidian languages are still underrepresenting themselves in the domain of computer linguistics, particularly when it comes to identifying abusive content (Mohan et al., 2023). This paper brings a highly effective approach to comment classification, abused as well as non-abused, for those languages with the use of a multilingual BERT model. This is an approach that does well on using labeled datasets and weighted loss functions for handling class imbalance (Shanmugavadivel et al., 2022). This paper tries to improve the detection of abuse in low-resource languages, giving important insights that will open avenues for further research on fighting online abuse in different linguistic domains (Rajalakshmi et al., 2023). This study is different from the previous ones because it is based on Tamil and Malayalam, which are languages with complicated patterns, hence making it hard to identify misuse (Subramanian et al., 2023). Multilingual BERT ensures that the model can efficiently handle code-mixed text and regional linguistic peculiarities (Ponnusamy et al., 2024). In addition, weighted loss functions and thorough preprocessing stages enhance the robustness of the proposed method (Shanmugavadivel et al., 2023).Limitations on the use of digital spaces include harassment against women and other disadvantage groups online. In the development of abuse classifiers for the ICON2023 Gendered Abuse Detection challenge, annotated Twitter datasets in English, Hindi, and Tamil were available. Combining two Ensemble Approach (EA), namely CNN and BiLSTM modeling contextual dependencies while CNN captures abusive language characteristics, an ensemble model is presented by the CNLP-NITS-PP team (Vetagiri et al., 2024). Comments on social media can shift the political and corporate climate overnight, affecting people and civilizations in ways that are impossible to ignore. But the same media also make it easier to launch targeted attacks on specific people or organizations. To identify hope speech and harmful remarks categorized as xenophobia, transphobia, homophobia, misogyny, misandry, and counter-speech, a shared job was implemented (Priyadharshini et al., 2022). Participants used a variety of Deep Learning (DL) and machine learning models on datasets that were either entirely Tamil or included a mix of Tamil and English codes. They then presented their findings and insights into how they used the datasets. The prevalence of social media calls for increased efforts toward the identification and classification of abusive remarks (Priyadharshini et al., 2023). In this process, there is an increasing need for Tamil and other low-resource Indic languages, which happen to provide greater obstacles (Reshma et al., 2023). The paper makes use of data augmentation methods like lexical substitution and back-translation along with multilingual transformer-based models for categorizing offensive comments posted on YouTube in Tamil. The Multilingual Representations for Indian Languages (MURIL) transformer model had the best performance with a 15-point improvement in macro F1-scores compared to baselines. This article discusses methods for efficient preparation of Tamil text for abuse detection.

3 Dataset Description

To classify abusive and non-abusive comments in Dravidian languages—Tamil and Malayalam—we used datasets consisting of development, training, and test data. Malayalam has 629 samples (Abusive-304, non-abusive-325); Tamil has 598 samples (Abusive-278, non-abusive-320); Two classifications comprise each datasets: Abusive and Non-Abusive. For every language, the test data consist of 558 Tamil and 629 Malayalam samples.

Language	Development Data	Training Data	Test Data
Tamil	598	2700	558
Malayalam	629	2933	629

Table 1: Details of the dataset used to classify abusive and non-abusive comments in Dravidian languages—Tamil and Malayalam.

4 Methodology

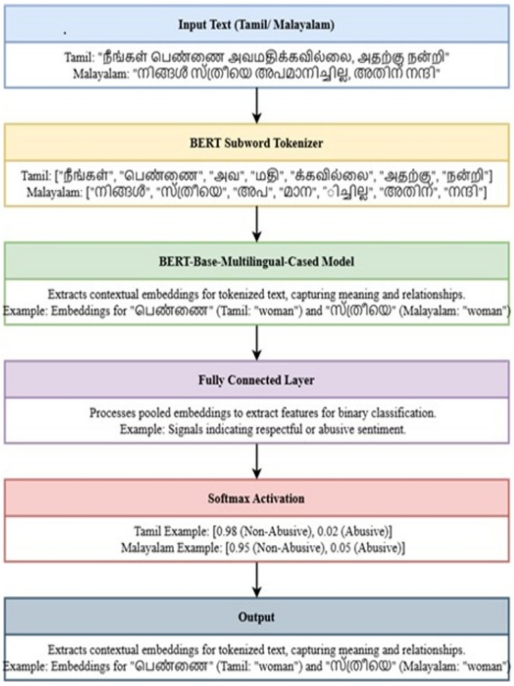
Input Text (Tamil/ Malayalam)
Tamil: "நீங்கள் பெண்ணை அவமதிக்கவில்லை, அதற்கு நன்றி"
Malayalam: "നിങ്ങൾ സ്ത്രീയെ അപമാനിച്ചില്ല, അതിന് നന്ദി"
BERT Subword Tokenizer
Tamil: ["நீங்கள்", "பெண்ணை", "அவ", "மதி", "க்கவில்லை", "அதற்கு", "நன்றி"]
Malayalam: ["നിങ്ങൾ", "സ്ത്രീയെ", "അപ", "മാന", "ിച്ചില്ല", "അതിന്", "നന്ദി"]
BERT-Base-Multilingual-Cased Model
Extracts contextual embeddings for tokenized text, capturing meaning and relationships.
Example: Embeddings for "பெண்ணை" (Tamil: "woman") and "സ്ത്രീയെ" (Malayalam: "woman")
Fully Connected Layer
Processes pooled embeddings to extract features for binary classification.
Example: Signals indicating respectful or abusive sentiment.
Softmax Activation
Tamil Example: [0.98 (Non-Abusive), 0.02 (Abusive)]
Malayalam Example: [0.95 (Non-Abusive), 0.05 (Abusive)]
Output
Extracts contextual embeddings for tokenized text, capturing meaning and relationships.
Example: Embeddings for "பெண்ணை" (Tamil: "woman") and "സ്ത്രീയെ" (Malayalam: "woman")

Figure 1: BERT Model Architecture

Figure 1 illustrates the model architecture that employs BERT for identifying offensive comments. Contextual embeddings are extracted from Tamil or Malayalam input text using BERT Subword Tokenizer for tokenization, and then it is processed by the BERT-base-multilingual-cased model. A Fully Connected Layer is applied for binary classification using these embeddings and a Softmax Activation function. The output layer categorizes the input as abusive or non-abusive, an efficient way of dealing with language variation and complexity.

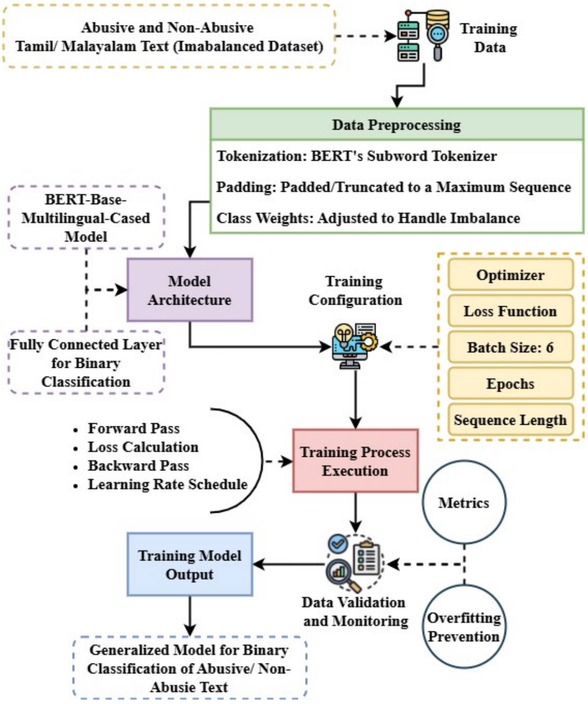
Abusive and Non-Abusive
Tamil/ Malayalam Text (Imabalanced Dataset)
Training Data
Data Preprocessing
Tokenization: BERT's Subword Tokenizer
Padding: Padded/Truncated to a Maximum Sequence
Class Weights: Adjusted to Handle Imbalance
BERT-Base-Multilingual-Cased Model
Model Architecture
Training Configuration
Optimizer
Loss Function
Batch Size: 6
Epochs
Sequence Length
Fully Connected Layer for Binary Classification
Forward Pass
Loss Calculation
Backward Pass
Learning Rate Schedule
Training Process Execution
Metrics
Training Model Output
Data Validation and Monitoring
Overfitting Prevention
Generalized Model for Binary Classification of Abusive/ Non-Abusie Text

Figure 2: Training process of BERT-Base-Multilingual-Cased Model

Figure 2 presents a workflow for abusive and non-abusive text classification in Tamil and Malayalam using a BERT-based model. The process begins with training data, which consists of an imbalanced dataset of abusive and non-abusive text. In the data pre-processing step, the BERT sub-word tokenizer is used for tokenization, while the sequences are padded or truncated to a fixed length. Additionally, class weights are adjusted to handle the imbalance of the data set. The model architecture is based on the BERT-Base Multilingual Cased Model, enhanced with a fully connected layer for binary classification. The training configuration specifies key hyper parameters such as the optimizer, loss function, batch size (6), number of epochs, and sequence length. During training execution, the model undergoes a forward pass, loss calculation, backward pass, and learning rate scheduling to optimize performance. Once trained, the model outputs a generalized classifier capable of distinguishing abusive and non-abusive text. Performance is assessed using evaluation metrics, and data validation and monitoring ensure model reliability. In addition, overfitting prevention techniques are applied to improve generalization. This structured approach uses deep learning and transfer learning techniques to improve abusive speech detection in low-resource Dravidian languages.

$$Xs{*}Ba[ki{-}ah] \rightarrow Ls[v{-}zw''] {+} Va[\partial\alpha{-}aqw''] \tag{1}$$

The equation represents the transformations of input token embeddings through the self-attention and optimization mechanisms of the BERT-Base Multilingual Cased model for abusive and non-abusive text classification. Here, Xs denotes the token embeddings, which are numerical vector representations of words in the input text. $Ba[ki - ah]$ refers to the attention-weighted representation of these token embeddings, capturing contextual relationships between words. The transformation yields $Ls[v - zw'']$, an intermediate feature representation obtained from the hidden layers, where zw'' represents refinements made by the self-attention mechanism. Additionally, $Va[\partial\alpha - aqw'']$ describes the gradient-based optimization process during model training, where $\partial\alpha$ represents parameter updates in backpropagation, and aqw'' indicates higher-order interactions in contextual learning. Together, these components refine word representations, ensuring that abusive language patterns are effectively captured while maintaining contextual integrity. This transformation ultimately enables the classification layer to accurately differentiate between abusive and non-abusive text.

5 Result and Discussion

The results indicate that the proposed BERT-based model outperforms existing methods across key evaluation metrics, making it highly effective for detecting abusive comments in Tamil and Malayalam.

5.1 Evaluation Metrics for Tamil

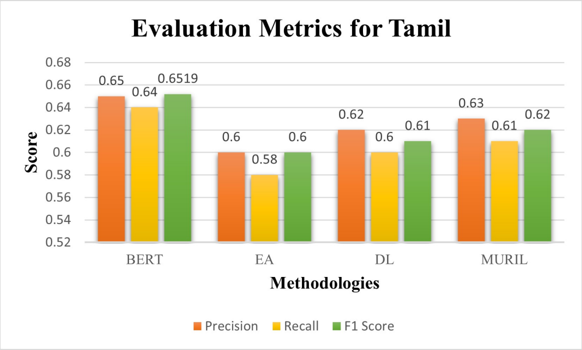

Figure 3: Evaluation Metrics for Tamil

The performance of BERT for the classification of abusive comments in Tamil achieved an F1 score of 0.6519, outperforming related methods such as EA (0.60), DL (0.61) and MURIL (0.62). This demonstrates BERT's ability to handle linguistic complexity and imbalanced datasets better. Preprocessing techniques, including script normalization and handling code-mixed text, significantly enhanced the classification quality. Challenges in capturing contextual nuances, such as sarcasm and implicit abuse, remain. Future work will explore advanced embedding techniques for better results.

$$Ds[\{li{-}an''\}] \rightarrow Ls[as{-}naq'']+Va[ds{-}iuwq''] \tag{2}$$

The equation describes the transformation of input data through the self-attention and optimization mechanisms in the BERT-Base Multilingual Cased model for abusive and non-abusive text classification. Here, $Ds[\{li{-}an''\}]$ represents the processed input embeddings, where li and an'' refer to specific token-level features, possibly modified by language embeddings or attention mechanisms. The right-hand side consists of two key components: $Ls[as - naq'']$, which denotes an intermediate feature representation extracted from the hidden layers, capturing the contextual relationships between words, and $Va[ds - iuwq'']$, which represents the gradient-based optimization process used for model fine-tuning. Here, ds may correspond to weight adjustments, while $iuwq''$ indicates advanced contextual refinements applied through backpropagation. This transformation ultimately helps the model better distinguish abusive text from non-abusive text by refining token relationships and optimizing the learned embeddings for classification.

5.2 Evaluation Metrics for Malayalam

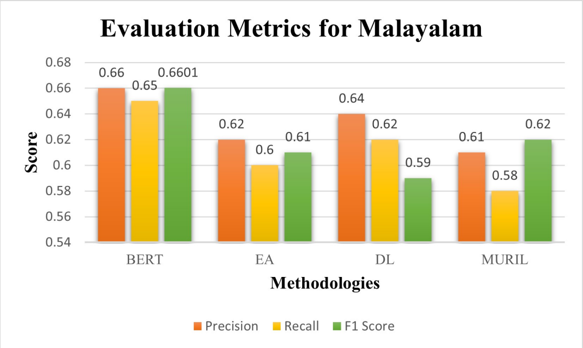

Figure 4: Evaluation Metrics for Malayalam

Although the F1-scores for EA, DL, and MURIL were all 0.66 for abusive comment classification in the Malayalam dataset, BERT achieved the highest F1-score of 0.6601. Using class-weighted loss functions that along with strong preprocessing greatly enhanced the performance of the model, therefore addressing the class imbalance and language variation. Although it beats other methods, the misclassification of more delicate cases of abusive language calls for improvement even if it exceeds other algorithms. Contextual embeddings and explainable artificial intelligence techniques help to identify abusive content in Malayalam even more.

$$p_{fvd}[k-anw''] \rightarrow Dsp[v-znq'']+Va[s-e6v''] \tag{3}$$

The equation represents a transformation in the BERT-Base Multilingual Cased model for abusive and non-abusive text classification, illustrating how token embeddings evolve through different layers of the model. Here, $p_{fvd}[k - anw'']$ denotes the initial token representation, where k represents token indices, and anw'' signifies preprocessed features, possibly modified by positional encodings and attention mechanisms. The right-hand side consists of two major components: $Dsp[v - znq'']$, which refers to an intermediate feature representation extracted from hidden layers after applying the self-attention mechanism, and $Va[s - e6v'']$, which describes the gradient-based optimization process in the fine-tuning stage of BERT. Here, s represents model parameters, while $e6v''$ indicates deeper refinements applied through backpropagation. Together, these transformations allow the model to accurately learn contextual relationships between words, thereby improving its ability to classify text as abusive or non-abusive.

6 Limitations

Only a limited sample of training data is used to train the model. Misclassification may result from difficulties identifying slang, sarcasm, and contextual meaning. Furthermore, the results of the study might not be entirely indicative of long-term trends or generalizable.

7 Conclusion

Using the BERT-Base Multilingual Cased model, this study demonstrates an efficient method for classifying abusive comments in Tamil and Malayalam.

Multilingual embeddings, combined with sophisticated preprocessing techniques and class-weighted loss functions, successfully addressed the challenges of language variation and class imbalance. The model achieved F1-scores of 0.6519 for Tamil and 0.6601 for Malayalam, outperforming the baseline models EA, DL, and MURIL. Significant efficiency improvements were achieved through preprocessing techniques such as script normalization and handling code-mixed text. Despite these advancements, challenges persist, particularly in the misclassifications of subtle expressions such as sarcasm and implicit abuse. Future work includes expanding the dataset with additional languages and diverse sources, developing multilingual embeddings tailored for low-resource languages, and incorporating explainable artificial intelligence to enhance interpretability. These advancements aim to improve abusive content detection systems, thereby enhancing online safety for vulnerable individuals.

References

B. R. Chakravarthi, R. Priyadharshini, S. Banerjee, M. B. Jagadeeshan, P. K. Kumaresan, R. Ponnusamy, 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. R. Mekapati, and B. R. Chakravarthi. 2023. A multimodal approach for hate and offensive content detection in Tamil: From corpus creation to model development. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

R. Ponnusamy, K. Pannerselvam, R. Saranya, P. K. Kumaresan, S. Thavareesan, S. Bhuvaneswari, and B. R. Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Siddhanth Shanmugavadivel, Kogilavani 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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, 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*.

R. Rajalakshmi, S. Selvaraj, and P. Vasudevan. 2023. Hottest: Hate and offensive content identification in Tamil using transformers and enhanced stemming. *Computer Speech & Language*, 78:101464.

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.

S. Reshma, B. Raghavan, and S. J. Nirmala. 2023. Mitigating abusive comment detection in Tamil text: A data augmentation approach with transformer model. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 460–465.

K. Shanmugavadivel, R. Chinnasamy, N. Subbarayan, A. Ganesan, D. Ravi, V. Palanikumar, and B. R. Chakravarthi. 2023. On finetuning adapter-based transformer models for classifying abusive social media Tamil comments.

K. Shanmugavadivel, S. U. Hegde, and P. K. Kumaresan. 2022. Overview of abusive comment detection in Tamil-acl 2022. In *DravidianLangTech 2022*, page 292.

M. Subramanian, K. Shanmugavadivel, N. Subbarayan, A. Ganesan, D. Ravi, V. Palanikumar, and B. R. Chakravarthi. 2023. On finetuning adapter-based transformer models for classifying abusive social media Tamil comments.

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*.