DLTCNITPY@DravidianLangTech 2025 Abusive Code-mixed Text Detection System Targeting Women for Tamil and Malayalam Languages using Deep Learning Technique

Habiba A , Aghila G

Department of Computer Science and Engineering,
National Institute of Technology Puducherry, India.

Abstract

The growing use of social communication platforms has seen women facing higher degrees of online violence than ever before. This paper presents how a deep learning abuse detection system can be applied to inappropriate text directed at women on social media. Because of the diversity of languages and the casual nature of online communication, coupled with the cultural diversity around the world, the detection of such content is often severely lacking. This research utilized Long Short-Term Memory (LSTM) for abuse text detection in Malayalam and Tamil languages. This model delivers 0.75, a high F1 score for Malayalam, and for Tamil, 0.72, achieving the desired balance of identifying abuse and non-abusive content and achieving high-performance rates. The designed model, based on the dataset provided in DravidianLangTech@NAACL2025 (shared task) comprising code-mixed abusive and non-abusive social media posts in Malayalam and Tamil, showcases a high propensity for detecting accuracy and indicates the likely success of deep learning-based models for abuse text detection in resource-constrained languages.

1 Introduction

The digital era has changed how people interact with each other and the rest of the world. Through social media, people can communicate with others who are oceans apart, share and experience events, and give opinions on matters openly. Social media has, however, advanced into a popular space that allows people to engage in passive yet harmful behaviours, most notably gendered abuse (De la Parra-Guerra et al., 2025). Some of the most common examples of these forms of abuse include the use of derogatory, threatening, or even demeaning language directed towards women and other targets (Gonzalez et al., 2025). This paper aims to comprehend the phenomena of aggressive Tamil and Malayalam language texts targeting women in social media by trying to find the patterns of the abuse and providing effective measures against them. Abusive language, particularly in social media, is communication intended to cause emotional or physical harm to others by using harmful or offensive verbal or textual content (Sinclair et al., 2025). Such verbal abuse may come in the form of name-calling, vulgar threats, and even sexual harassment. It is also aimed towards people who fall into specific categories, like women, ethnic minorities, or people from lower economic status. When this abuse is directed at women, it lacks compassion at its core, and this type of internet abuse is very damaging because it inflicts grave psychological, social and economic harm while reinforcing abuse of gender discrimination. The reason that understanding such abusive texts in Tamil and Malayalam is essential is due to a rise in online abuse within the South Indian linguistic paradigm. Among Indian languages, Tamil and Malayalam are among the most widely spoken. And their speakers form a vast and diverse population on social media. However, there have been no systematic studies on the use of abusive language in these languages and its impact on the dominant gender, particularly women. This understanding is necessary to form protective measures against unrelenting online abuse towards women in Tamil and Malayalam languages. In addition, these regional languages need to be studied to formulate more effective strategies focusing on prevention. The abuse is commonly used as a means to silence women, intimidate them, or disparage their voices in public debates, especially for women who speak about matters dealing with politics, gender, or social justice issues. For instance, women journalists, activists, and other women in powerful and visible public positions experience extreme forms of online abuse across all platforms, even in English, Spanish or Arabic. The verbal abuse is aimed at asserting supremacy over women and lowering

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 640–645
May 3, 2025 ©2025 Association for Computational Linguistics

their self-esteem to the point where they are afraid to speak in public spheres (Albladi et al., 2025). There is ample literature that documents the issue of online abuse of women in different languages (Priyadharshini et al., 2022b, 2023b). Still, much less research has been done on how such abuse exists in Dravidian languages such as Tamil and Malayalam. The construction of identity is based on the language, and those regional languages that shape identity also come with cultural baggage that defines the abuse. Traditional gender norms, social discrimination, and particular dialects may determine the form that violence and abuse will take in Tamil and Malayalam. Hence, it deliberately illustrates the need to design an abusive text detection strategy for Malayalam and Tamil, incorporating informal language structures on various social media platforms.

2 Related Works

Existing models on abusive text detection have been successful in languages such as English, but their feasibility for languages like Malayalam and Tamil with less annotated data has been largely unexplored. These languages are uniquely different from English as both are Dravidian languages (Chakravarthi et al., 2021; Priyadharshini et al., 2023a), making the application of existing models impractical. Moreover, the challenge intensifies because large-scale annotated data for these languages are not available to the public. The informal use of language on social networks adds to the problem. For example, the frequent use of slang, abbreviations, code-mixing, and code-switching creates more problems for the existing text classification models. The following section outlines key findings from recent studies addressing this issue. Machine learning has been widely adopted to classify abusive language over the past few years. Support Vector machines (SVM), Random forests, and Naive Bayes classifiers have been used in (Mahmud et al., 2024; Thavareesan and Mahesan, 2019) to categorize abusive language using pre-defined features of the text. Nevertheless, a lot of them (Aljero and Dimililer, 2020; HaCohen-Kerner and Uzan, 2021) depend on excessive feature engineering, such as the formation of words or n-grams for hate speech detection. However, the emergence of deep learning has enormously impacted text classification processes and detecting abuses in text. RNNs and their later developments, Long Short-Term Memory (LSTM) networks, are used in (Zhang, 2024; Al-Qerem et al., 2024). Gated Recurrent Units (GRUs) are highly efficient in comprehending sequential information and perfect for text. These models can understand the relation between words and phrases in contexts, which helps identify abuses in context-sensitive language. More recent approaches, such as BERT, which is a transformer model, have been successful in (Tarun et al., 2024) over earlier methods of abusive text detection due to the model's ability to comprehend higher levels of representation semantics. The problem of abusive text detection in languages with relatively little data is complex. The presence of low-quality annotated datasets for languages such as Malayalam and Tamil makes it harder for models to be trained and to attain performance standards. In addition, the informal language that includes slang, code-mixing, code-switching, and dialect makes it even more challenging to detect abuse consistently and is also highlighted in (Priyadharshini et al., 2022a; Subramanian et al., 2024). Efforts carried out in languages with limited resources have shown that there is a need to develop a specific approach for such languages that considers a language's unique features.

3 Methodology

This section describes the detection of abusive text in the Malayalam and Tamil languages using an LSTM. The methodology comprises the following steps: dataset gathering, feature extraction, model training, and evaluation. Figure1 illustrates the general pipeline used for the text detection system. The user-generated content from YouTube, a social media platform where women are abused, is the source from which a dataset of abusive and non-abusive text is compiled. This data set is gathered from (Rajiakodi et al., 2025), the shared task DravidianLangTech@NAACL2025, and contains text samples in Malayalam and Tamil. The steps utilized for the detection system are elaborated in Algorithm 1.

3.1 Datasets

Here, we focus on two specifically code-mixed datasets: Malayalam code-mixed, which contains text that has both Malayalam and English words, and Tamil code-mixed text, which contains text that has both Tamil and English. Here, we focus on two specifically code-mixed datasets: Malayalam code-

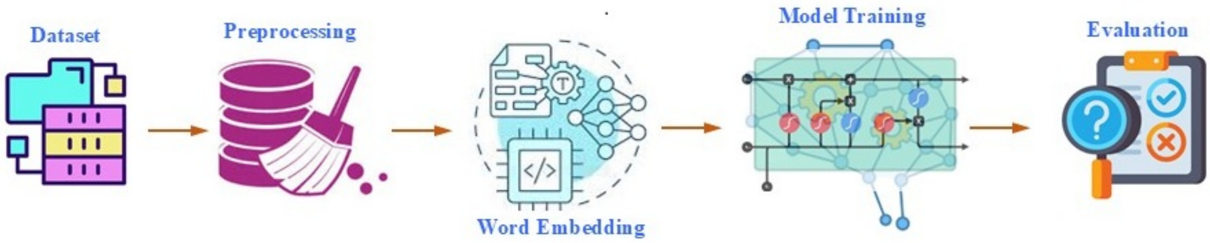

Figure 1: Text Detection System Pipeline

mixed, which contains text that has both Malayalam and English words, and Tamil code-mixed text, which contains text that has both Tamil and English. Table 1 describes the detailed statistics of both datasets. Both include text data categorized as abusive or non-abusive.

	Training Data		Validation Data	
	Abusive	Non-Abusive	Abusive	Non-Abusive
T	1366 (49%)	1424 (51%)	278 (46%)	320 (54%)
M	1531 (52%)	1402 (48%)	303 (48%)	326 (52%)

Table 1: Statistics of the dataset used for Tamil (T) and Malayalam (M).

3.2 Preprocessing

Before being used for analysis, text data goes through various cleaning and standardization steps, an essential step in almost every Natural Language Processing task. First, all text is translated into lowercase format. While this addresses the need for the first essential step for cleaning and standardizing text data, it also ensures no capitalization issues occur. Then, the unwanted spaces are removed from the dataset. As a result of checking whether the data has been prepared adequately for analysis, it serves as the basis for developing an efficient text classification model. Text data in the model undergoes tokenization, transforming each word into a unique number. The models scarcely use an extensive vocabulary and don't entertain less common words. The tokenized text is also padded so that all input data has the same shape.This preprocessed data is fed into the model for training.

3.3 Our Approach

The model used in this study is an LSTM, a specific RNN structure capable of working with sequences like string text. An architecture of this type includes an embedding layer, LSTM layer, and dense output layer. Word indices are mapped to fixed-sized dense vectors by the embedding layer. In contrast, the spatial dropout layer, which is set over the embedding layer, decreases overfitting for the model by setting a proportion of the embedding inputs to blank. The LSTM layer maintains the sequential relationships among the words. To avoid overfitting, dropout is used with both the input and recurrent connections. The dense output layer has two units enhanced with a sigmoid activation function for 2 class target variables. The model is compiled with Adam as the optimiser and binary cross-entropy loss, performing well for binary classification problems. During training, the model accuracy throughout evaluation is maintained. The model is trained for five epochs with a batch size of 64. The hyperparameter settings used to train the model are shown in the table 2. The validation data assesses the model's performance on the new test datasets that the model has not seen after every epoch. The metric F1-score is used to check the model's performance.

Parameters	**Value**
Embed Units	EMBEDDING_DIM = 100
Hidden Units (LSTM)	100
Dropout	0.2
Optimizer	Adam
Batch Size	64
Loss	Binary Crossentropy
Epochs	5
Activation	Sigmoid

Table 2: Hyperparameter Table

Algorithm 1 Text Classification for Abusive Language Detection using LSTM

Input: Dataset D with Text Sequences S

Output: Categories [$Abusive$ or $Non-Abusive$].

The input sequences are tokenized.

$$Tokenized(S) = [t_1, t_2, t_3, ..., t_n]$$

where t_i is the token for word w_i.

Using an embedding layer, each token t_i is mapped to a dense vector representation.

$$Embedding(t_i) = (e_{i1}, e_{i2}, e_{i3}, ..., e_{id})$$

where e_i represents the embedding vector for token t_i.

The embeddings are passed through the LSTM, and the output is a sequence of hidden state vectors:

$$h_i = LSTM(e_i, h_{i-1})$$

where h_i is the hidden state at time step i, and h_{i-1} is the previous hidden state.

The final hidden state h_n is used for classification.

The dense layer produces a vector y_{logits} for the final prediction:

$$y_{logits} = Wh_n$$

where W is the weight matrix and h_n is the final hidden state.

The output probabilities are calculated:

$$p(y) = [p(0), p(1)]$$

where $p(0)$ is the "Non-Abusive" class probability, and $p(1)$ is the "Abusive" class probability.

The binary cross-entropy loss function is used to measure the discrepancy between the predicted probabilities:

$$BC\ Loss = -\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i)$$

where N is the number of samples in the batch, y_i is the actual label for the i-th sample, and p_i is the predicted probability for the i-th sample.

4 Results

The model improved the accuracy of the Tamil code mixed dataset throughout the training epochs. The accuracy result is 78.48%, and the F1 score is 0.7207 for the Tamil test set. The model placed 18$^{\text{th}}$, with an impressive quantitative score. The model also improved significantly during the training epochs for the Malayalam code mixed dataset. The accuracy achieved was 73.50, and an F1 score of 0.7571 from the Malayalam dataset, showing balanced classification results. It is impressive that the model ranked 1 for Malayalam in the shared task.The table 3 shows the results achieved through our approach. The model handled the Tamil and Malayalam datasets well, making it the best model for this category.

Team Name	Language	F1	Rank
Habiba A, Aghila G (This work)	Malayalam	0.7571	1
	Tamil	0.7207	18

Table 3: Result Achieved

5 Discussions

Although our model performed well, we recognize the weaknesses of the LSTM architecture. In future, we have a strategy to resolve developmental possibilities, such as looking into more complex neural networks, incorporating more factors such as sentiment, and taking advantage of more significant and varied datasets.

6 Conclusion

This paper outlines a methodology wherein deep neural networks detect abusive texts in multiple languages, especially those that lack sufficient training data, like Malayalam and Tamil. Using RNNs and their ability to detect patterns, we aim to create an accurate model for abuse text detection in women. The findings, as appreciated, need an elaboration on concepts like an informal conversation, the use of idioms, and the absence of sufficient labelled datasets. The model is trained for each particular language separately to accommodate their diverse alterations in abusive language.

7 Error Analysis

While the LSTM model used on Tamil and Malayalam hate speech shows excellent accuracy, especially with true positives, it faces challenges with false positives and other classed errors within a supposedly balanced dataset. This particular behaviour needs much attention and improvement on the model's discrimination capabilities. Evaluation of the model's performance through comprehensive validation and test set analyses is vital for determining performance generalizability. Some of the outlined tactical changes include changing the degree of false positive identification and incorporating large language models for optimization.

8 Limitations

The pertinent issues of this analysis are the inadequacy of the annotated dataset for both Malayalam and Tamil—more significant datasets aid model effectiveness and generalizability. The preprocessing and classification of such texts are complicated due to the lack of their formal quality. And indeed, language, as used in social media, constructed with informal terms, regional variations, and shortened expressions, is highly challenging. Most of these phrases lack grammatical structure, which hinders traditional models from determining the sentiment of the text. Both these languages, Malayalam and Tamil, are classified under the same Dravidian language family yet differ in morphology, syntax, semantics and other unique factors.

Ethics Statement

We maintain compliance with ACL guidelines in participating DravidianLangTech@NAACL2025 shared task, as well as our commitment to ethical research practices. No ethical issues or conflicts of interest emerged throughout the duration of this research.

References

Ahmad Al-Qerem, Mohammed Raja, Sameh Taqatqa, and Mutaz Rsmi Abu Sara. 2024. Utilizing deep learning models (rnn, lstm, cnn-lstm, and bi-lstm) for arabic text classification. In *Artificial Intelligence-Augmented Digital Twins: Transforming Industrial Operations for Innovation and Sustainability*, pages 287–301. Springer.

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892.

Mona Khalifa A. Aljero and Nazife Dimililer. 2020. Hate speech detection using genetic programming. *2020 International Conference on Advanced Science and Engineering (ICOASE)*, pages 1–5.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. Mccrae. 2021. Dravidiancodemix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56:765 – 806.

AC De la Parra-Guerra, J Truyol-Padilla, CA García-Alzate, and F Fuentes-Gandara. 2025. Gender-based violence as a barrier to women rights towards socio-environmental sustainability. *Global Journal of Environmental Science and Management*, 11(1):343–364.

Alejandra Gonzalez, James K Haws, Nuha Alshabani, Caron Zlotnick, and Dawn M Johnson. 2025. Cyber abuse and posttraumatic stress disorder among racially diverse women who have resided in domestic violence shelters: A longitudinal approach. *Psychological Trauma: Theory, Research, Practice, and Policy*.

Yaakov HaCohen-Kerner and Moshe Uzan. 2021. Detecting offensive language in english hindi and marathi using classical supervised machine learning methods and word/char n-grams. In *FIRE (Working Notes)*, pages 501–507.

Tanjim Mahmud, Tahmina Akter, Mohammad Kamal Uddin, Mohammad Tarek Aziz, Mohammad Shahadat Hossain, and Karl Andersson. 2024. Machine learning techniques for identifying child abusive texts in online platforms. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022a. 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, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022b. 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, S Malliga, Subalalitha Cn, SV Kogilavani, B Premjith, Abirami Murugappan, and Prasanna Kumar

Kumaresan. 2023a. 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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023b. 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.

Gary Sinclair, Colm Kearns, Katie Liston, Daniel Kilvington, Jack Black, Mark Doidge, Thomas Fletcher, and Theo Lynn. 2025. Online abuse, emotion work and sports journalism. *Journalism Studies*, 26(1):101–119.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

VG Tarun, Ramkumar Sivasakthivel, Gobinath Ramar, Manikandan Rajagopal, and G Sivaraman. 2024. Exploring bert and bi-lstm for toxic comment classification: A comparative analysis. In *2024 Second International Conference on Data Science and Information System (ICDSIS)*, pages 1–6. IEEE.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on industrial and information systems (ICIIS)*, pages 320–325. IEEE.

Hongmin Zhang. 2024. Research on text classification based on lstm-cnn. In *Proceeding of the 2024 5th International Conference on Computer Science and Management Technology*, pages 277–282.