KEC_TECH_TITANS@DravidianLangTech 2025: Abusive Text Detection in Tamil and Malayalam Social Media Comments Using Machine Learning

Malliga Subramanian[1], Kogilavani S V[1], Deepiga P[1], Dharshini S[1], Ananthakumar S[1], Praveenkumar C[1]

[1]*Kongu Engineering College, Erode, Tamil Nadu, India*

Abstract

Social media platforms have become a breeding ground for hostility and toxicity, with abusive language targeting women becoming a widespread issue. This paper addresses the detection of abusive content in Tamil and Malayalam social media comments using machine learning models. We experimented with GRU, LSTM, Bidirectional LSTM, CNN, FastText, and XGBoost models, evaluating their performance on a code-mixed dataset of Tamil and Malayalam comments collected from YouTube. Our findings demonstrate that the the FastText and CNN models yielded the best performance among the evaluated classifiers, achievF1 scoresores of 0.73 each. This study contributes to ongoing research on abusive text detection for under-resourced languages and highlights the need for robust, scalable solutions to combat online toxicity.

1 Introduction

The rise of social media has revolutionized how individuals communicate, share opinions, and engage with global communities. However, this unprecedented connectivity comes at the cost of an alarming increase in abusive language, particularly targeting women. Abusive language not only perpetuates gender inequality, but also has severe psychological and social consequences. Addressing this issue requires efficient tools to detect and mitigate this content effectively.

Previous works on abusive text detection have predominantly focused on English, leaving low-resource languages like Tamil and Malayalam underexplored. Moreover, the code-mixed nature of these languages further complicates the task, as traditional monolingual models fail to handle linguistic complexities inherent in such data. Building on the growing body of research on offensive language detection, this study proposes the application of machine learning models for classifying Tamil and Malayalam social media comments as abusive or non-abusive.

2 Literature Survey

The rise of social networks has required automated methods to detect and mitigate offensive content (Blair, 2003). While fostering global communication, social platforms have also become hubs for harmful language targeting individuals and groups. Advances in natural language processing (NLP) have enabled sophisticated systems to classify abusive language, even in multilingual and code-mixed contexts (Lee and Kim, 2015). However, detecting nuanced, context-dependent abuse remains challenging due to its subjective nature and linguistic variations (Obadimu, 2020).

Early studies demonstrated the effectiveness of machine learning models like Support Vector Machines (SVMs) and Naive Bayes, which relied on handcrafted features such as n-grams and TF-IDF. Deep learning models like CNNs and RNNs further improved classification by capturing contextual and sequential text patterns (T. De Smedt, 2018; Waseem and Hovy, 2016). Ribeiro et al. (M. H. Ribeiro and Jr, 2018) analyzed hateful behavior on Twitter using machine learning, while Kshirsagar et al. (P. Mishra and Shutova, 2018) highlighted the role of predictive embeddings in enhancing hate speech detection.

More recently, transformer-based models such as BERT and RoBERTa have set new benchmarks in offensive language detection by leveraging large-scale pretraining and fine-tuning (J. Mitrović and Granitzer, 2019; Fortuna and Nunes, 2019). These models effectively capture complex linguistic structures, making them ideal for tackling abusive language detection.

Despite these advancements, their application to low-resource and code-mixed languages, like Tamil and Malayalam, remains underexplored (C. Nobata,

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 346–350

2016). Code-mixed text presents challenges such as irregular grammar, mixed scripts, and context-switching, which existing models trained on high-resource languages struggle to address (Schmidt and Wiegand, 2018). Bridging this gap is essential for developing inclusive tools that curb online abuse across diverse linguistic communities.

3 Materials and Methods

3.1 Task Description

This study classifies Tamil and Malayalam social media comments as either abusive or non-abusive. The dataset consists of YouTube comments annotated with binary labels:

- Abusive
- Non-Abusive

3.2 Dataset

The dataset includes Tamil and Malayalam code-mixed comments from YouTube, annotated based on content. It consists of 5,000 comments, with an average sentence length of 1. Figure 1 presents sample texts.

Text	Label
உங்கள் முயற்சி வெற்றியடைய வாழ்த்துகள்!	Non-Abusive
നിന്നക് രോബിനെ കല്യാണം കഴിക്കണ്ടന്ന് ശരിക്കും ഉള്ള സത്യം.	Non-Abusive
நீ வெறும் வேலைக்கு ஒத்திகை இல்லாத நபர்!	Abusive
നവ്യയുടേ കയ്യിന് കീഴിലല്ലോ. ഇവാക്ക് അതോടെ എറിയില്ലേ	Abusive

Figure 1: Sample training texts from the dataset are shown below.

3.3 Preprocessing and Feature Extraction

Preprocessing was essential for effective classification and involved:

3.3.1 Text Cleaning

Noise such as punctuation, special characters, and emojis was removed. Emojis were converted into textual descriptions to retain sentiment (J. Salminen, 2020).

3.3.2 Tokenization

Text was split into individual tokens, allowing models to analyze semantic patterns and relationships (Gao and Huang, 2020).

3.3.3 Feature Extraction

TF-IDF vectorization assigned weights to words based on frequency, ensuring focus on informative features (H. Mubarak and Magdy, 2017; A. Vidgen, 2020). This transformation structured the data for machine learning models (Agrawal and Awekar, 2018).

3.4 Models

We evaluated various models for abusive content detection:

- **GRU**: A recurrent neural network (RNN) capturing text dependencies efficiently.
- **LSTM**: Addresses the vanishing gradient problem, effectively handling long-range dependencies.
- **Bidirectional LSTM**: Enhances context understanding by processing sequences in both directions.
- **CNN**: Extracts n-gram-like features, making it computationally efficient for classification.
- **FastText**: Embedding-based model averaging word vectors for classification.
- **XGBoost**: Gradient boosting framework leveraging decision trees for structured data classification.

4 Results and Discussion

4.1 Performance Metrics

The models were evaluated using **Accuracy, Precision, Recall, and F1-Score**, as summarized in Figure 2.

These commonly used evaluation metrics are defined as follows:

- **Accuracy**: The proportion of correctly classified texts:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- **Recall (Sensitivity)**: The proportion of correctly classified texts in a class:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

- **Precision (Positive Predictive Value)**: The proportion of correct predictions per class:

Model	Precision	Recall	F1-Score	Accuracy
GRU	0.69	0.69	0.69	69%
LSTM	0.69	0.69	0.69	69%
Bidirectional LSTM	0.26	0.50	0.34	52%
CNN	0.73	0.73	0.73	73%
FastText	0.73	0.73	0.73	73%
XGBoost	0.68	0.67	0.67	67%

Figure 2: **Performance Metrics Table.**

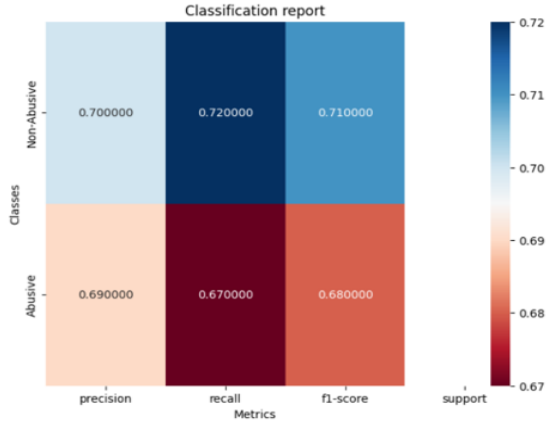
$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

- **F1-Score**: The harmonic mean of Precision and Recall:

$$\text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

4.2 Model Performance Analysis

The classification performance was analyzed using these metrics. The detailed reports for selected models are shown below (see Figures 3, 4, 5, and 6).



(a) GRU Model

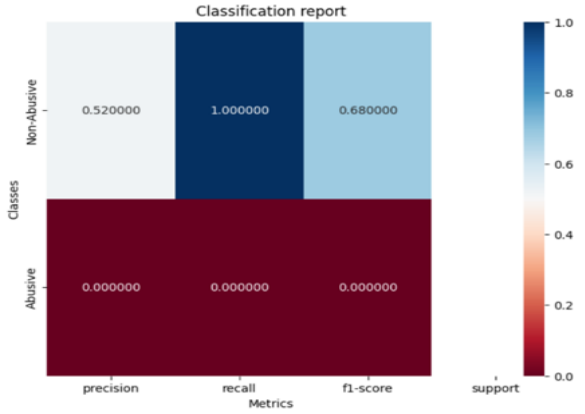
Figure 3: **GRU Model Performance**

5 Error Analysis

To better understand the challenges in detecting abusive content, we performed both qualitative and quantitative error analysis.

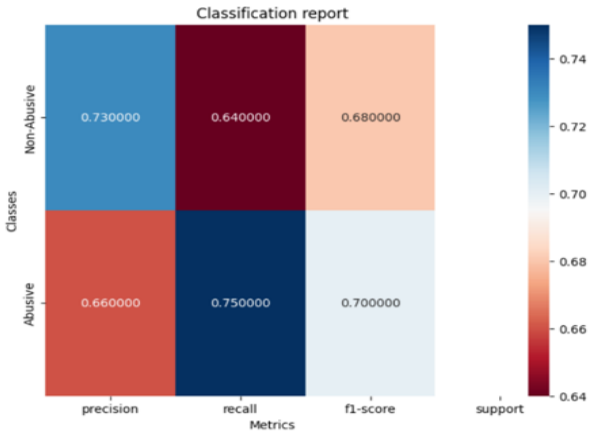
5.1 Qualitative Analysis

We manually inspected misclassified examples to identify patterns. Some key observations include:



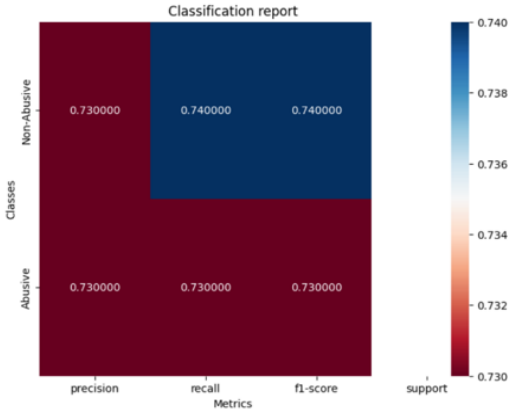
(c) Bidirectional LSTM

Figure 4: **Bidirectional LSTM Model Performance**



(b) LSTM Model

Figure 5: **LSTM Model Performance**



(d) CNN Model

Figure 6: **CNN Model Performance**

- Code-mixed comments with informal spelling variations were often misclassified.
- Sarcasm and implicit abuse were challenging for the models to detect accurately.
- Certain abusive words were misclassified due to their different contextual meanings.

5.2 Quantitative Analysis

We examined key misclassification trends:

- The CNN and FastText models performed well but misclassified some non-abusive comments as abusive.
- GRU and LSTM models struggled with long-text dependencies, leading to errors.
- Class imbalance affected the F1-score, causing a bias toward the majority class.

This analysis highlights the need for improved preprocessing techniques and context-aware abuse detection.

5.3 Hyperparameter Settings

The hyperparameters used for training the models are summarized in Table 1.

Hyperparameter	Value
Learning Rate	0.001
Batch Size	32
Dropout Rate	0.3
Number of Epochs	10
Optimizer	Adam
Loss Function	Cross-Entropy Loss

Table 1: Hyperparameter settings used for training models.

6 Conclusion

This study conducts a comparative evaluation of machine learning models for detecting abusive content in Tamil and Malayalam social media comments. The results reveal that CNN and FastText models achieved superior performance, with each attaining an F1-Score of 0.73. These findings highlight the effectiveness of these models in addressing the complexities of code-mixed and low-resource language datasets, where traditional methods often struggle. Despite this success, there remains considerable scope for improvement(Schmidt and Wiegand, 2018; Zhang and Luo, 2018). Future work will explore cutting-edge transformer-based architectures like BERT, RoBERTa, and multilingual models, which have shown significant promise in other language processing tasks. Additionally, advanced feature representation techniques, (Chakravarthi et al., 2025) such as contextual embeddings and hybrid feature extraction methods, will be investigated to enhance the models' capability to capture nuanced and context-dependent abusive language more effectively.

7 Limitations

While our approach demonstrates promising results in detecting abusive and sentiment-based text in low-resource languages, several limitations remain:

- **Data Imbalance:** The dataset contains an uneven distribution of classes, which may lead to biased predictions, especially for underrepresented labels.
- **Code-Mixed Challenges:** Handling code-mixed text remains complex due to variations in spelling, grammar, and transliteration across languages.
- **Generalization:** The trained models may not generalize well to unseen datasets or different social media platforms due to variations in language usage.
- **Computational Constraints:** Transformer-based models require significant computational resources, making deployment on low-end devices challenging.
- **Contextual Limitations:** Certain comments require deeper contextual understanding, which current models may struggle to interpret accurately.

Project Repository

The full source code for this project is available on GitHub: GitHub Repository - Deepikagowtham

References

et al. A. Vidgen. 2020. Challenges and frontiers in abusive content detection. *arXiv preprint arXiv:2010.07395*.

S. Agrawal and A. Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Proceedings of the European Conference on Information Retrieval*.

J. Blair. 2003. New breed of bullies torment their peers on the internet. *Education Week*, 22:6.

et al. C. Nobata. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview of the shared task on political multiclass sentiment analysis of tamil x(twitter) comments: Dravidianlangtech@naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

A. Fortuna and S. Nunes. 2019. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 52(4).

T. Gao and M. Huang. 2020. Detecting online hate speech using context-aware models. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*.

K. Darwish H. Mubarak and W. Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*.

B. Birkeneder J. Mitrović and M. Granitzer. 2019. nlpup at semeval-2019 task 6: A deep neural language model for offensive language detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

et al. J. Salminen. 2020. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. *ACM Transactions on Social Computing*.

S.-H. Lee and H.-W. Kim. 2015. Why people post benevolent and malicious comments online. *Communications of the ACM*, 58:74–79.

Y. A. Santos V. A. Almeida M. H. Ribeiro, P. H. Calais and W. Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth International AAAI Conference on Web and Social Media*.

A. M. Obadimu. 2020. *Assessing the Role of Social Media Platforms in the Propagation of Toxicity*. Ph.D. thesis, University of Arkansas at Little Rock.

H. Yannakoudakis P. Mishra and E. Shutova. 2018. Neural character-based composition models for abuse detection. *arXiv preprint arXiv:1809.00378*.

J. Schmidt and A. Wiegand. 2018. A survey on hate speech detection using natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*.

E. Kotzé L. Saoud M. Gwóźdź G. De Pauw et al. T. De Smedt, S. Jaki. 2018. Multilingual cross-domain perspectives on online hate speech. *arXiv preprint arXiv:1809.03944*.

Z. Waseem and D. Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection. In *NAACL Student Research Workshop*, pages 88–93.

M. Zhang and Y. Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *arXiv preprint arXiv:1803.03662*.