TAMILGOODBADTXT@DravidianLangTech 2026:A Multilingual Transformer-Based Approach for Abusive Language Identification in Tamil Social Media

K Varalakshmi[1], **B Bharathi**[2]
[1] St. Joseph's Institute of Technology, Tamil Nadu, India
[2]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India
varaluckky.2@gmail.com, bharathib@ssn.edu.in

Abstract

It is difficult to detect abusive language, particularly in social networks for low-resource languages like Tamil. Spelling errors, informal expressions and code-mixing make it even more challenging to read text from social media. The current work proposes a multilingual transformer-based approach to detect abusive content in Tamil text. A pretrained XLM-RoBERTa model is used to learn contextual and semantic representations from the input text. This is a general pipeline comprising preprocessing, tokenization, and binary classification (abusive / non-abusive). Experiments are performed on Tamil social media datasets that have abusive and non-abusive data. The results reveal that multilingual transformer models achieve good performance in low-resource scenarios. The proposed model attains an F1-score of 78.64%, which shows the potential of using cross-lingual pretrained models for the detection of abusive Tamil language.

Keywords: Abusive language detection; Tamil NLP; multilingual transformers; XLM-RoBERTa; Low-resource languages.

1 Introduction

Social network sites have seen unprecedented user-generated content, which has seen individuals express their views, feelings, and thoughts on the Internet under the freedom of expression. This has contributed to the rapid spread of abusive, violent, and hate speech as much as it has promoted digital communication, and this is a very big threat to man and society at large. Therefore, the automatic detection of the existence of abusive language has become a significant concern in the natural language processing field that has significance in offering content regulation and in offering a safer Internet experience. Tamil is a low-resource Dravidian language, and the fact that the abusive text is identified in Tamil social media presents certain challenges. Online Tamil is very informal and normally contains a variation of spelling and dialect and is highly code-mixed with English. These characteristics render the conventional rule-based systems as well as the conventional machine learning techniques to be less efficient, in part due to the fact that there are no huge annotated data sets. As such, it would not be an easy task to create effective Tamil abusive language detectors. The more recent progress in deep learning, particularly transformer-based language models, has demonstrated itself to be able to perform state-of-the-art in a large variety of text classification tasks. Rich multilingual cross-lingual contextual representations Multilingual transformer models are trainable on large-scale multilingual corpora with rich cross-lingual contextual representations. In this case, low-resource languages, such as the XLM-RoBERTa model specifically, are especially suitable to transfer linguistic information of the high-resource languages and improve learning of representation and classification performance. This study addresses the challenge of detecting abusive language in the text of Tamil social media through multilingual transformer-based models. The proposed system trains a multilingual model that is already trained on Tamil-labeled data to identify the texts as abusive or not. Such semantic relationships and contextual clues that are involved with the linguistic variation and informal style of writing that is evident in Tamil social media can be dealt with successfully by the system. The results of the experiment demonstrate that the use of multilingual transformer models is efficient in the condition of the absence of the data and language barriers, and they may be considered a viable option in the situation with the abusive language detection in the low-resource languages(i.e Tamil)

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 392–396

2 Literature Survey

The detection of abusive language has also received a lot of focus particularly in low-resource and multilingual environment. The initial methods used were based on the traditional machine learning and deep neural networks like CNNs and RNNs to detect hate speech in social media textual data (Zhang et al., 2018). Later on, their performance was enhanced by transformer-based models such as BERT which is able to capture contextual semantics and thus enhance its performance over others (Mozafari et al., 2019). OffensEval shared task focused on the detection of types of offensive contents and targets (Zampieri et al., 2019).

As multilingual transformers emerged, studies were expanded to low resource languages. Rajalakshmi et al. showed the efficiency of Tamil multilingual models to detect abusive comments (Rajalakshmi et al., 2022). Recent works at DravidianLangTech workshops included Tamil and Malayalam on pretrained transformers, such as detecting women-targeted abusive text in a series of works on the topic, including (Thirumoorthy et al., 2025; Al Nahian et al., 2025; Bade et al., 2025).

Multilingual toxicity detection: Multilingual hybrid models based on XLM-RoBERTa with CNN and BiLSTM have demonstrated encouraging outcomes (Singhal et al., 2025). Moreover, cross-lingual generalization was also enhanced by large language model-based approaches (Usman et al., 2025). The article by Nalini et al. offers an overview of the latest innovations (Nalini et al., 2024). The previous research article by Gamback and Sikdar, as well as Davidson et al. (Gambäck and Sikdar, 2017; Davidson et al., 2017), indicated that CNNs have been proven to be effective and that differentiating between hate speech and offensive language proves to be challenging.

Transformer-based models also enhanced the detection of abusive language. Das et al. (Das et al., 2021) and Mishra et al. (**?**) demonstrated the superiority of BERT, mBERT, and XLM-R when it comes to a multilingual and low-resource setting. Studies of Dravidian languages in recent times still utilize pretrained transformers of Tamil and Malayalam (Thirumoorthy et al., 2025; Al Nahian et al., 2025).

The results of the article by (Rajiakodi et al., 2026) reveal that the issues that were encountered include morphological richness and the necessity of powerful classification models in the Tamil language. In spite of this, such problems as linguistic diversity, code-mixing, and annotated data are still present. Out of these challenges, this paper introduces a transformer-based multilingual model that is fine-tuned on the detection of Tamil abusive language.

3 Proposed Work

In this work, the input data consists of Tamil text and a fine-tuned multilingual transformer model is used to capture the contextual and semantic representations to detect abusive language. The system is transformer-based Where raw Tamil text is provided as input and the outputs are binary classification. Model agnostic, enabling a variety of pretrained transformer backbones to be used. It includes pre-processing of text, generation of contextual embedding and classification. The complete architecture of the proposed system is illustrated in Figure 1.

3.1 Input Dataset Representation

A corpus of Tamil social media posts in CSV format was fed into the system as input data. There are 3,652 samples in the training set, categorized into two classes: Abusive and Non-Abusive. The test data consists of 913 examples, which are used for testing. Each example sentence is a Tamil sentence without any metadata. To ensure linguistic diversity, no manual processing is involved in the process.

3.2 Text Encoding and Tokenization

Each of the input sentences is segmented into a series of subword tokens using a model-specific tokenizer.Each sentence in the input is split into a sequence of subword tokens, using a model-specific tokenizer. This process creates input token IDs, attention masks and pads to a fixed sequence length. The method proved to be effective in capturing the variations in the morphology and the code-mixing expressions in the Tamil social media text.

3.3 Transformer-Based Feature Extraction

A pretrained transformer encoder is used to encode the text into deep contextual representations. These are the multilingual transformer models that are tested:

- XLM-RoBERTa Base
- IndicBERT
- MuRIL

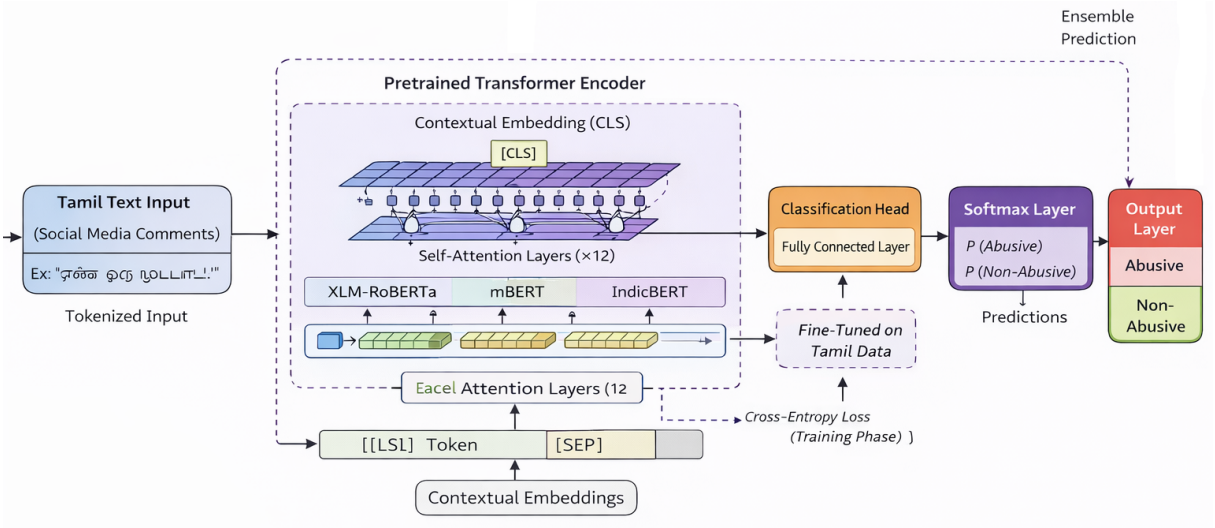

Figure 1: Proposed transformer-based architecture for abusive language detection in Tamil text

Multi-head self-attention is adopted in these models to model long-range dependency and the semantic context. The pretrained representations are fine-tuned to better suit the abusive language detection task.

3.4 Classification Layer

The contextual embeddings produced by the transformer encoder are fed to a classification head comprising a fully connected neural network with a SoftMax activation function. The model is predicting probabilities for the two classes Abusive and Non-Abusive. The system is trained with cross entropy loss function and optimized with the AdamW optimizer.

3.5 Output Generation

In inference, the trained model is used to predict the class label for the unseen test data. The expected results are then turned into the final class labels that are evaluated with standard performance measures.

4 Dataset

The data set this work uses is for detection of abusive language in the Tamil social media comments. It is split into two sets: a training set and a test set in order to perform supervised learning and evaluation. It has 3652 sentences in the training set and 913 sentences in the test set, totaling 4565 sentences.

The dataset is provided as part of a shared task on Tamil abusive language detection. All training samples are manually annotated as **Abusive** or **Non-Abusive**. The dataset reflects real-world social media characteristics, including informal language, spelling variations, colloquial expressions, emojis, and code-mixing with English.

Table 1: Dataset Statistics

Dataset Split	Samples	Annotation
Training	3,652	Abusive / Non-Abusive
Test	913	Unlabeled
Total	**4,565**	–

The training data is imbalanced, with **Non-Abusive** as the majority class and **Abusive** as the minority class, reflecting real-world distribution patterns.

The source code GitHub link is available here: [1].

5 Experimental Set Up And Results Analysis

The model training was carried out on a GPU environment with the aim of efficient convergence and computational performance improvement. A standard preprocessing pipeline was applied to the input text to ensure uniformity in the data, which includes normalization, tokenization with transformer-based tokenizers, padding, truncation to the maximum sequence length, and creation of attention masks needed by the model. The objective function for training the classification problem (abusive vs not abusive) was binary cross-entropy

[1] https://github.com/Varalakshmi2793/TAMILGOODBADTXT.git

loss. Optimization was carried out with Adam optimiser with the proper learning rate schedule to guarantee a stable convergence. The model was trained end-to-end to acquire multilingual contextual representations and semantics and the syntactic properties of Tamil social media texts.

Model trained was tested against 913 unseen samples to assess the proposed system. Standard classification metrics such as Accuracy, Precision, Recall, Macro F1-Score, and Weighted F1-Score were used to measure the performance. Accuracy is the overall correctness of predictions. Precision is the ratio of correctly predicted samples that are abusive to all samples predicted as abusive. Recall is the model's ability to accurately detect real abusive content. The F1-Score gives both the Precision and the Recall a balance. Macro F1-Score treats both classes equally and Weighted F1-Score considers the class imbalance.

The quantitative performance of the proposed model (RUN 1) is presented in Table 2.

Table 2: Performance Evaluation Metrics

Metric	Value
Accuracy	0.7864
Precision	0.7881
Recall	0.7876
Macro F1-Score	0.7864
Weighted F1-Score	0.7800

It is found that, the model has an overall accuracy of 78.64%. The accuracy of 78.81% implies that the model is able to identify abusive content with relatively low amount of false positive and the recall value of 78.76% shows that maximum number of abusive instances are identified. The Macro F1-Score of 78.64% shows that the model performs well on both classes, and the Weighted F1-Score of 78.00% takes into account the impact of class imbalance. The model does not show strong class bias, with the close values of precision and recall. Additional information about the classification behavior is presented in the confusion matrix shown in Figure 2.

Based on the classification results in the 913 test samples, the model achieved True Negatives (TN) = 359, False Positives (FP) = 98, False Negatives (FN) = 97 and True Positives (TP) = 360. The results show that a good percentage of the samples were classified correctly, far more than misclassified samples.

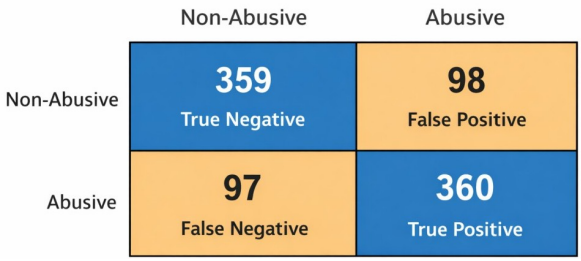

Figure 2: Confusion matrix of the proposed abusive language detection model

The near equivalence of the false positive and false negative rates suggest that the model is not heavily skewed towards either of the two classes of abusive or non-abusive. Ambiguous language, sarcasm, implicit abusive language, informal language and code-mixed Tamil-English text often used on social media platforms are some of the reasons for misclassification. In general, the multilingual transformer based model is effective for detecting abusive language in low resource setting of Tamil language and has good generalization capability.

6 Conclusion

The work proposed a multilingual transformer based approach for Tamil abusive language detection. The model was trained on annotated Tamil text and evaluated on 913 test samples. The proposed system achieved 78.64% accuracy with balanced value of precision, recall and F1 score. The results show that transformer-based contextual embeddings are capable of capturing the semantic and syntactic properties of Tamil text, and hence they can effectively detect abusive language in low-resource settings. The analysis of the confusion matrix also shows a stable classification performance without a significant bias to any class. Future work could focus on incorporating bigger domain specific datasets, handling code-mixed and sarcastic texts more efficiently and exploring ensemble transformer architectures to improve the performance further.

References

Abdullah Al Nahian, Mst Rafia Islam, Azmine Toushik Wasi, and Md Manjurul Ahsan. 2025. Nlpopsciol@ dravidianlangtech 2025: Classification of abusive tamil and malayalam text targeting women using pretrained models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 38–45.

Girma Yohannis Bade, Zahra Ahani, Olga Kolesnikova, José Luis Oropeza, and Grigori Sidorov. 2025. Gs_dravidianlangtech@ 2025: Women targeted abusive texts detection on social media. *arXiv preprint arXiv:2504.02863*.

Mithun Das, Somnath Banerjee, and Punyajoy Saha. 2021. Abusive and threatening language detection in urdu using boosting based and bert based models: A comparative approach. *arXiv preprint arXiv:2111.14830*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International conference on complex networks and their applications*, pages 928–940. Springer.

C Nalini, R Shanthakumari, Y Agashia Maria, T Janarthanan, and M Manibharathi. 2024. Advancements in offensive language detection: A comprehensive review and experimental analysis. *Journal of Information Assurance & Security*, 19(4).

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022. Dlrg@ dravidianlangtech-acl2022: Abusive comment detection in tamil using multilingual transformer models. In *Proceedings of the second workshop on speech and language technologies for Dravidian languages*, pages 207–213.

Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Rajalakshmi R., Kathiravan Pannerselvam, Bhuvaneswari Sivagnanam, Jananayagan V, Charmathi Rajkumar, R Ramesh Kannan, and Bharathi Raja Chakravarthi. 2026. From Comments to Harm: A Findings Report on Abusive Tamil Text Targeting Women on Social Media Shared Task- DravidianLangTech@ACL 2026. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Nikita Singhal, Avadhesh Yadav, Ankush Ankush, Giriraj Singh, and Ronak Kumar. 2025. Leveraging xlm-roberta with cnn and bilstm for hinglish toxicity detection. *Journal of Communications Software and Systems*, 21(4):394–403.

Shanmitha Thirumoorthy, Thenmozhi Durairaj, and Ratnavel Rajalakshmi. 2025. Hydrangea@ dravidianlantech2025: Abusive language identification from tamil and malayalam text using transformer models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 580–584.

Muhammad Usman, Muhammad Ahmad, Grigori Sidorov, Irina Gelbukh, and Rolando Quintero Tellez. 2025. A large language model-based approach for multilingual hate speech detection on social media. *Computers*, 14(7):279.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.