Lannisters@DravidianLangTech 2026: A Comparative and Ablation Study of Multilingual Transformers for Gender-Targeted Abuse Detection in Tamil Social Media Platforms

Kalaivani K S[1], **Jaisanth K**[1], **Nandhini B**[1],
[1]Department of Artificial Intelligence and Data Science,
Kongu Engineering College, Perundurai, Erode
{kalaivani.cse, jaisanthk.23aid, nandhinib.23aid}@kongu.edu

Abstract

The prevalence of the use of the Tamil language on social media has heightened the need to address the issue of online harassment of women. As a result, there is a heightened need to develop a system to automatically identify abusive content in the Tamil language to promote a safe online communication platform. This paper presents a model to identify abusive content using a binary classification model to identify Abusive and Non-Abusive content. In this work, we experimented with several multilingual transformer models including DistilBERT, mBERT, and XLM-RoBERTa. From the experiments, it was observed that the XLM-RoBERTa model performed better than the others, achieving a validation accuracy of 91.17% and a validation macro F1 score of 0.8865. In this paper, ablation experiments are conducted to show that structured preprocessing, balancing the minority class, and tuning the hyperparameters contribute to the model's performance.

1 Introduction

With the rise of social media platforms, the way Tamil-speaking users communicate online has changed quite noticeably. Though social media enables the sharing of ideas and the maintenance of interpersonal contacts, it also offers scope for the dissemination of abusive and gender-related content targeting women. In this context, online harassment occurs through the use of direct insult, aggressive undertones, sarcasm, and other forms of subtle communication. It may also reflect the social inequality among the genders. It is not feasible to depend only on human evaluators for the task. It is therefore necessary to explore the possibilities offered by machines. Abusive content detection in the Tamil script is a task with a unique set of challenges. The script is morphologically complex and not fully covered by the existing literature. Online social media messages also often contain spelling and grammar mistakes, words and expressions from different dialects, and words with unconventional spellings. Abusive content is also complex and context-dependent. It is not feasible to adopt a rule-based and keyword-based approach for the task.

Recent studies indicate that multilingual transformer-based language models show promising results for detecting abusive content, especially in low-resource and cross-lingual settings. It is also possible to capture the meaning of words with the help of the transformer architecture. Transformer-based models such as BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and XLM-RoBERTa (Conneau et al., 2020) have significantly improved performance in multilingual natural language processing tasks. In this study, we aim to examine how different multilingual transformer architectures perform in detecting abusive content written in the Tamil language.

2 Related Work

Previous research on Tamil abusive and offensive language classification using machine learning techniques has demonstrated the effectiveness of contextualized word embeddings. For example, Kalaivani et al. (2021) compared traditional machine learning methods, deep learning approaches, and multilingual transformer models for identifying offensive Tamil-English code-mixed comments. Their results showed that Multilingual BERT achieved the best macro F1 score, highlighting the effectiveness of contextualized embeddings for code-mixed language classification.

In another study, (Patankar et al., 2022) explored abusive comment detection in Tamil and Tamil-English social media text using ensemble models, recurrent neural networks, and transformer-based approaches. Their find-

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 289–293

ings showed that MuRIL and XLM-RoBERTa achieved strong performance for abusive language classification in low-resource settings. Multilingual models designed specifically for Indian languages, such as MuRIL, have shown promising performance in cross-lingual natural language processing tasks Kakwani et al. (2021). Recent studies on abusive language detection in Dravidian languages have shown that multilingual transformer-based approaches are effective for handling low-resource and code-mixed text. Shared tasks such as DravidianLangTech have encouraged the development of robust abusive language detection systems for Tamil social media data.

A new shared task on the detection of offensive language in Dravidian languages was proposed by Chakravarthi et al. (2021). An analysis of abusive comments in Tamil directed towards women on social media was carried out by Sivagnanam et al. (2026), and the results indicated that women were being targeted based on gender. Hate speech detection has been widely studied in natural language processing.

Hate speech detection has been widely studied in natural language processing, with surveys by Davidson et al. (2017), Fortuna and Nunes (2018), and Zhang et al. (2022) providing comprehensive overviews of detection techniques and datasets.

3 Dataset

The dataset includes abusive language directed towards women in the form of social media comments in the Tamil language. It is a binary classification problem with two classes: Abusive and Non-Abusive. The training set contains 3,652 instances, while the test set contains 913 instances.

The dataset exhibits class imbalance, with the Non-Abusive class containing more instances than the Abusive class. To solve this problem, data augmentation was performed on the Abusive class, increasing the instances in this class. This resulted in a total of 7,190 instances.

The dataset was split into training and validation sets in an 80:20 ratio, giving 5,752 instances in the training set and 1,438 in the validation set. Due to class imbalance, the evaluation metric used was the macro-average metric to assess model performance. Several shared tasks and benchmark datasets have been created to promote the study of abusive language detection in Dravidian languages (Chakravarthi et al., 2021).

4 Methodology

4.1 Data Preprocessing and Augmentation

Significant noise is present in social media text, such as HTML tags, Unicode characters, and repetitive punctuation and spelling varieties. In order to reduce these noise-related problems, a preprocessing pipeline is applied before training the model, including HTML entity decoding and HTML tag removal.

The overall workflow of the proposed abusive content detection system is illustrated in Figure 1.

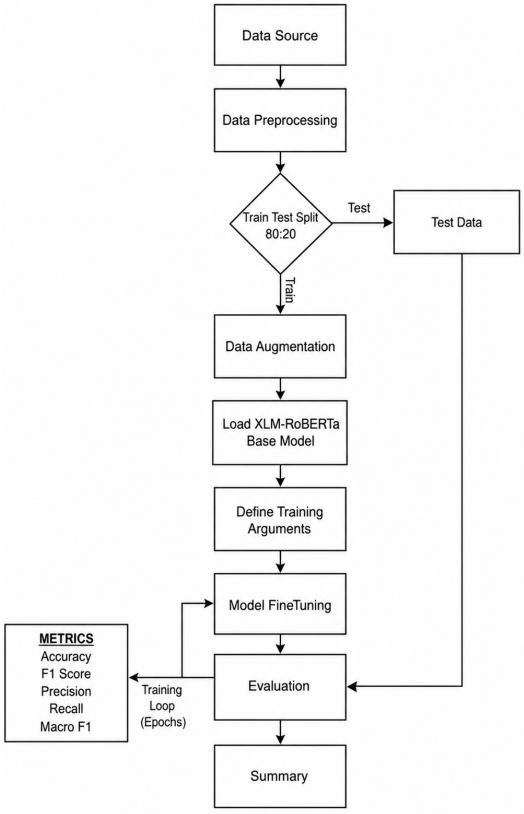

Figure 1: Proposed system workflow

Unicode normalization was applied to normalize the Unicode character set. This includes the removal of zero-width characters and the conversion of Tamil numerals to regular numerals. URLs, emails, and user mentions were removed. However, the removal of hashtags was avoided since they play an essential role in determining the lexical content. Obfuscated abusive words were replaced with a predefined word. Repeated punctuation marks were suppressed, and repeated words were smoothed. Emojis were avoided since they are helpful in determining the sentiment of

the content. Classes with small instance counts were balanced using techniques like Classes with small instance counts were balanced using techniques such as random deletion, random word swapping, and back-translation. Back-translation was performed using English as the intermediate language, and approximately equal proportions of augmented samples were generated from each augmentation technique.

4.2 Model Architecture

Three different transformer model architectures were tested: DistilBERT, mBERT, and XLM-RoBERTa. For each model architecture, the model was fine-tuned on the binary sequence classification task with the same preprocessing and training settings.

The XLM-RoBERTa model architecture is based on a transformer model with multiple stacked layers in the encoder. For the classification task, the token representing the input sequence is passed through a dropout layer with a dropout probability (p = 0.1), and then passed through a linear classification layer. The maximum sequence length is 256 tokens. Cross-entropy class weighting is used to compute the loss.

5 Experiments

5.1 Model Comparison

For selecting the best architecture, we conducted a side-by-side test for DistilBERT, mBERT, and XLM-RoBERTa under the same conditions. The validation accuracy results are summarized in Table 1. The results indicate that XLM-RoBERTa achieves the highest performance among the evaluated models, demonstrating its stronger capability to capture the morphological characteristics of the Tamil language and code-mixed patterns.

Table 1: Validation accuracy of evaluated models

Model	Validation Accuracy (%)
DistilBERT	87.4
mBERT	89.2
XLM-RoBERTa	**91.17**

5.2 Training Strategy

XLM-RoBERTa was trained for six epochs with a learning rate of 2e-5, batch size of 16, weight decay of 0.01, and a smoothing level of 0.1. In order to handle the class imbalance, class weights were used, and early stopping was based on the macro F1 score obtained during validation. In the next step, we trained again on the entire data set to maximize exposure before making the final test predictions.

Hyperparameter tuning was performed by experimenting with multiple learning rates, batch sizes, and epoch settings. The final configuration was selected based on the highest validation macro F1 score.

All experiments were conducted using PyTorch and HuggingFace Transformers on an NVIDIA Tesla T4 GPU.

6 Results and Analysis

6.1 Validation Performance

The validation performance of the optimized XLM-RoBERTa model is summarized below.

Table 2: Validation performance of the optimized XLM-RoBERTa model

Metric	Score
Accuracy	91.17%
Weighted F1	0.9119
Macro F1	0.8865
F1 (Abusive)	0.94
F1 (Non-Abusive)	0.83

The reported results correspond to the local validation set. Final predictions for the official shared-task evaluation were generated separately using the organizer-provided blind test set.

6.2 Training Dynamics

The training dynamics of the model are illustrated in Figure 2, which presents both the training and validation loss as well as accuracy trends across epochs.

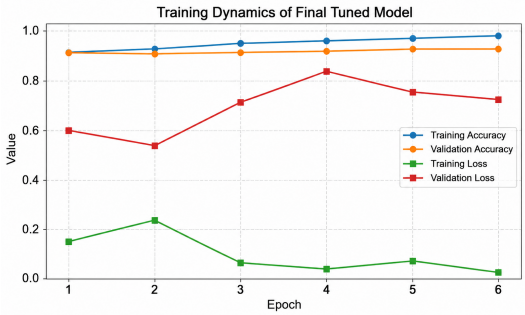

Figure 2: Training and validation accuracy and loss curves

The curves indicate stable convergence during training without significant overfitting. The validation accuracy improves steadily while the loss decreases across epochs, demonstrating effective learning behavior of the model.

6.3 Error and Confidence Analysis

The classification performance of the model is further analyzed using the confusion matrix shown in Figure 3.

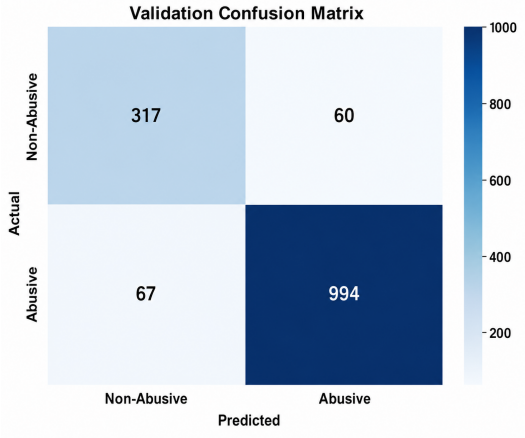

Figure 3: Validation confusion matrix

The confusion matrix indicates that the model achieves strong recall for abusive content detection. However, some misclassifications occur in comments containing sarcasm, implicit abuse, dialectal variations, and code-mixed Tamil-English expressions. Non-abusive comments containing emotionally intense language were occasionally misclassified as abusive. These observations suggest that contextual and pragmatic understanding remains challenging for multilingual transformer models. For example, certain sarcastic Tamil comments containing indirect abusive expressions were incorrectly classified due to contextual ambiguity.

6.4 Ablation Study

The ablation study considered four different scenarios:

- No preprocessing or augmentation
- Preprocessing only
- Preprocessing with data augmentation
- Preprocessing and augmentation followed by hyperparameter tuning

The performance improvements across these configurations are illustrated in Figure 4.

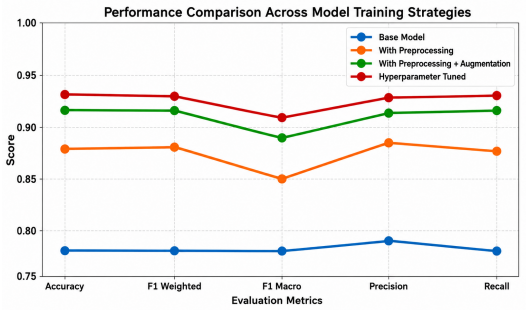

Figure 4: Ablation performance stages

Table 3: Ablation study performance across different configurations

Configuration	Acc.	Macro F1
Base Model	0.78	0.78
Preprocessing	0.88	0.85
Preprocessing + Aug.	0.92	0.89
Hyperparameter Tuned	**0.93**	**0.91**

The ablation study demonstrates that preprocessing significantly reduces noise in social media text, while augmentation improves minority class representation. Hyperparameter tuning further stabilizes convergence and improves macro F1 performance.

7 Conclusion

This paper undertook an in-depth assessment of the effectiveness of multilingual transformer-based models in abusive language detection in the Tamil language. Comparative experiments show the effectiveness of the XLM-RoBERTa model over DistilBERT and mBERT. The XLM-RoBERTa model showed an accuracy of 91.17% on the validation set and a macro F1 of 0.8865. Future work will focus on improving contextual

understanding of sarcasm, implicit abuse, and dialectal variations in Tamil social media text.

8 Limitations

The dataset used in this paper remains limited in scale. This may limit the generalization of the results to other online Tamil contexts. The augmentation strategy may not capture the full range of variation in natural language. The model may have relatively poor performance in Non-Abusive instances and may have difficulty perceiving sarcasm and bias. In addition, this approach may have to be adapted to other social media sites for it to be applicable.

9 Ethical Considerations

Abuse detection using this model may have the potential to enable protection of users. However, false positives may limit freedom of expression. False negatives may allow abusive content to escape detection. Bias in diverse dialects and communities may have to be closely monitored. This model may have to be used in conjunction with human operators.

10 Code Availability

Code and implementation details are publicly available at:
https://github.com/Nandhinibalu05/SocialMediaTamilAbuse

References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John Philip McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in tamil, malayalam, and kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, and Vishrav Chaudhary. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Paula Fortuna and Sergio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4).

Shubham Kakwani, Anoop Kunchukuttan, and Satish Golla. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

S. Kalaivani, R. K. S. Bharathi, and P. Priyadharshini. 2021. Offensive language identification in tamil-english code-mixed social media text. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE)*.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. Optimize_Prime@DravidianLangTech-ACL2022: Abusive comment detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–239. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert: a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of the NeurIPS Workshop*.

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

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2022. Hate speech detection: A literature review. *ACM Computing Surveys*, 54(5).