DPR@DravidianLangTech 2026: Transformer-Based Abusive Content Detection for Tamil Text Targeting Women on Social Media

Diya Prakash[1] **and Praveen Kumar S**[1] **and Ranjith Kumar R**[1]
Siranjeevi Rajamanickam[2] and Balasubramanian Palani[1] and Jobin Jose[1]
[1]Indian Institute of Information Technology Kottayam, India
[2]Government Polytechnic College, Trichy, India
diyaprakash2205@gmail.com, spraveenkumar2205@gmail.com,
ranjith23bcd30@iiitkottayam.ac.in, rajasiranjeevi@gmail.com,
pbala@iiitkottayam.ac.in, jobin@iiitkottayam.ac.in

Abstract

The fast-growing number of content in Tamil in social media has led to increasing abusive and gender-directed hate speech in online platforms. Detecting abusive content written in Tamil is relatively difficult owing to the complex morphological structure of Tamil language, its dialects, transliteration, and contextualized usage. In this study, the use of transformer-based pretrained language models in detecting abusive content in Tamil was explored. Five transformer-based models—mBERT, MuRIL, XLM-RoBERTa, IndicBERT, and Tamil-BERT—were fine-tuned and tested using DravidianLangTech 2026 shared task dataset. The experimental results show that the best-performing model was Tamil-BERT with an accuracy rate of 80.72% owing to Tamil-specific pretraining and better morphological analysis capabilities. Our system ranks 5th at the leaderboard of the DravidianLangTech 2026 shared task challenge. The source code and fine-tuned models are open-source and publicly accessible.[1]

1 Introduction

Online Social Networks foster connectivity but also spread harmful, gender-biased abuse that causes significant personal and professional damage. Detecting this in Tamil is uniquely difficult due to its complex morphology, dialectal variations, and the use of transliteration or sarcasm. Since many forms of harm are context-specific, simple profanity checklists are inadequate for accurate identification. Therefore, advanced models are needed to analyze deeper semantic and syntactic language properties to effectively filter abusive content.

The primary contribution of this work lies in the comparative evaluation and analysis of transformer-based pretrained language models for Tamil abusive content detection in a low-resource setting. With a reference to the existing literature, we believe that transformer-based, pretrained language models such as mBERT, MuRIL, XLM-RoBERTa, IndicBERT and Tamil-BERT models can be effectively employed to classify resource scarce languages like Tamil. We believe that the transformer models will be superior to the traditional feature-engineering based approaches, as they can learn a variety of contextual features from large amounts of pretrained data and they can be beneficial for the abusive language detection task in Tamil language.

The key contributions of this work include:

- This research examines the performance of different transformer based architectures on Tamil language.
- This study analyzes the impact of multilingual versus Tamil-specific pretraining, along with WordPiece and SentencePiece tokenization strategies, on abusive language classification performance.
- We are contributing to the development of an automated moderation system for Tamil digital spaces as part of our research on hate speech detection for Dravidian languages.

2 Literature Survey

Detecting abusive language is a relatively new and rapidly evolving field of research, driven largely by the growth of social media platforms, where millions of users are increasingly spending their time online. Initially, researchers used traditional machine learning approaches such as Logistic Regression and Support Vector Machines (SVM) in conjunction with manually curated features (Kalaivani and Thenmozhi, 2024; Subramanian et al., 2022). These features were often n-grams, or TF-IDF representations of words. However, these approaches have been shown to only offer a limited solution to the classification of abusive language and do not deal with some of the complexities, especially the

[1] https://github.com/praveen-2205/DPR-DravidianLangTech2026

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 242–247

context-dependent, nuanced and implicit aspects of the language.

Recent studies have explored transformer-based and deep learning approaches for abusive and offensive language detection in Tamil (Mohan et al., 2025; Pokrywka and Jassem, 2024; Abishek and Govindapillai, 2024). Shared task systems from DravidianLangTech workshops have further demonstrated the effectiveness of transformer-based architectures for abusive Tamil text targeting women on social media (Rajiakodi et al., 2025; Bhattacharyya et al., 2025; Sangeetham and Bedi, 2025; Kumaresan and Karnati, 2025; Alam and Rahman, 2025; Thayasivam and Wijesuriya, 2025). Multimodal approaches have also been explored for misogyny and meme detection in Tamil and Malayalam (Hegde and Hande, 2025; Mallik et al., 2025; Ponnusamy et al., 2024).

The Transformer based architectures were introduced to the text classification tasks and achieved great success. The contextualized embeddings learned by models such as BERT and its multilingual versions, improve the performance of the classification models. The multilingual BERT (mBERT) show great cross-lingual generalization ability. Low-resource languages benefit from cross-lingual models. But for Indian languages, models such as MuRIL and IndicBERT are designed to handle the unique characteristics of Indic scripts which have high inflection and large syntactic variation. Tamil-BERT is a pretrained language model for the Tamil language. It learns the patterns present in the Tamil corpora and as a result, it enhances the language-specific constructions present in the Tamil language.

3 Methodology

This study focuses on the binary classification of Tamil YouTube comments into abusive and non-abusive categories using transformer-based architectures. The methodology involves fine-tuning pretrained transformer models for the classification task.

The architecture of the transformer-based classification model is shown in Figure 1, where Tamil comments are first tokenized and then passed through the transformer encoder to obtain contextual embeddings. The embedding corresponding to the classification token is subsequently used by a fully connected layer to predict whether the comment is abusive or non-abusive.

3.1 Problem Statement

Tamil abusive content detection is formulated as a binary text classification problem. The task is to classify a given Tamil YouTube comment into one of the two classes: *Abusive content against women* or *Not Abusive*. The classifier should be able to identify both the explicit and implicit content of abusive language, including: obscene language, derogatory terms, sarcasm, coded language.

3.2 Transformer-Based Embedding

Transformer-based models are well suited for Tamil abusive language detection because they capture contextual meaning more effectively than traditional keyword-based approaches. Tamil social media conversations frequently contain slang, sarcasm, transliterated text, spelling variations, and code-mixed expressions, making abusive intent highly context-dependent.

In this work, we evaluate five pretrained transformer models: mBERT, MuRIL, XLM-RoBERTa, IndicBERT, and Tamil-BERT. While mBERT and XLM-RoBERTa provide multilingual contextual representations, MuRIL and IndicBERT are specifically designed for Indian languages and code-mixed text. Tamil-BERT, pretrained exclusively on Tamil corpora, better captures Tamil-specific morphology and linguistic patterns.

All models use subword tokenization techniques such as WordPiece or SentencePiece, which improve handling of rare vocabulary, transliterated words, and morphological variations commonly observed in Tamil social media text. The pretrained models were fine-tuned for binary sequence classification using the contextualized representation of the `[CLS]` token.

4 Experimental Setup

This section describes the dataset usage, model training configuration, and evaluation procedure followed for Tamil abusive content detection.

4.1 Dataset

The dataset used in this study was released as part of the DravidianLangTech 2026 shared task on abusive content detection in Tamil. The dataset consists of Tamil YouTube comments manually annotated into two classes: *Abusive against Women (0)* and *Non-Abusive (1)*.

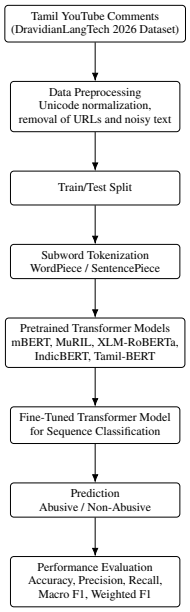

Figure 1: Overall workflow of the proposed transformer-based Tamil abusive content detection system.

Table 1: Dataset statistics for Tamil abusive content detection task.

Split	Abusive (0)	Non-Abusive (1)	Total
Training	1769	1883	3652
Test	441	472	913
Total	2210	2355	4565

Evaluation Metrics: We evaluate model performance using standard classification metrics: Accuracy, Precision, Recall, and F1-score. In addition to Macro F1-score, we also report Weighted F1-score to account for the slight class imbalance in the dataset.

4.2 Preprocessing

Minimal preprocessing was performed to preserve the informal nature of Tamil social media text, including slang, transliterated words, misspellings, and code-mixed expressions. The pretrained tokenizer corresponding to each transformer model was used for subword tokenization, and sequences were padded or truncated to a maximum length of 128 tokens.

4.3 Training Configuration

We fine-tuned five pretrained transformer models and adapted them for the downstream task of binary sequence classification. The logits are passed through a softmax activation function to compute the final class probabilities. The training was performed using supervised learning with cross-entropy loss. The models were trained for a fixed number of epochs with early stopping to prevent overfitting. We used the AdamW optimizer along with an appropriate learning rate schedule.

The models were trained with a maximum sequence length of 128, batch size of 32, learning rate of 2e-5, and 3 training epochs.

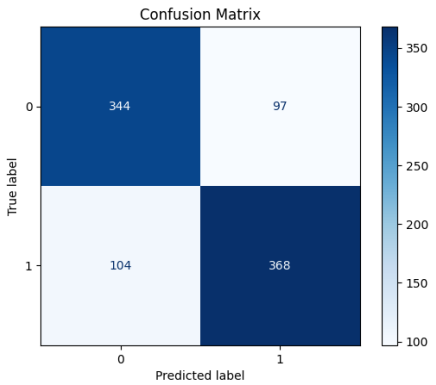

Figure 2: Confusion matrix for Tamil-BERT Model.

Table 2: Performance comparison of transformer models for Tamil abusive content detection.

Model	Accuracy	Precision	Recall	Macro F1	Weighted F1
mBERT	0.7711	0.7635	0.8072	0.7848	0.7591
MuRIL	0.7919	0.8294	0.7521	0.7889	0.7908
XLM-RoBERTa	0.7338	0.7338	0.7338	0.7338	0.7854
IndicBERT	0.7049	0.6901	0.7778	0.7313	0.6966
Tamil-BERT	**0.8072**	**0.8072**	**0.8076**	**0.8072**	**0.7851**

5 Results and Discussion

5.1 Model Performance

Table 2 presents the performance comparison of the five transformer models on the test set. Among the submitted systems in the DravidianLangTech 2026 shared task, our best-performing Tamil-BERT model achieved 5th rank on the official leaderboard, demonstrating the effectiveness of language-specific pretrained transformer models for Tamil abusive content detection.

5.2 Error Analysis

To better understand the behavior of different transformer models, we qualitatively analyzed representative examples from the test dataset. The analysis reveals that Tamil-specific and Indic-language pretrained models generally perform better on informal Tamil social media conversations containing slang, sarcasm, transliterated expressions, and contextual abuse.

Table 3 presents representative cases where some transformer models produced incorrect predictions. Tamil-BERT consistently demonstrated stronger contextual understanding due to its Tamil-specific pretraining, while multilingual models occasionally

Table 3: Representative error analysis examples.

Input Sentence	Model Prediction	Observation
"Ayyo sugandhida nambara thangalen da semma naatu kattaiya irukku"	MuRIL → Abusive	Informal slang and highly colloquial expressions were incorrectly associated with abusive intent despite the conversational tone being non-abusive.
"Rendu perum paththini maadhiri pesuraaluga"	XLM-RoBERTa → Abusive	Sarcastic wording and implicit conversational tone introduced ambiguity, leading the model to incorrectly classify the sentence as abusive.
"Appa saami moolai kalangiduchchi enakku... naan poi gp muthu video paathu relax pannikka poren"	IndicBERT → Non-Abusive	The model struggled to capture subtle emotional and contextual cues expressed through informal conversational Tamil and mixed emotional tone.

struggled with colloquial Tamil expressions and implicit conversational meaning.

The document discusses challenges faced in abusive language detection in Tamil social media text through three examples. The first example emphasizes the difficulties with highly colloquial Tamil slang, resulting in MuRIL misclassifying non-abusive conversational slang as abusive, while Tamil-BERT succeeded due to its Tamil-specific pretraining. The second example points to the challenges posed by sarcasm; XLM-RoBERTa misinterpreted sarcastic language as offensive, indicating that multilingual models may depend too much on lexical cues, ignoring conversational nuances. The third example reveals pitfalls in understanding emotional expressions, where IndicBERT erroneously classified a frustrated expression as abusive due to its inability to grasp nuanced emotional context and informal conversational structure.

Overall, the qualitative analysis suggests that both pretraining corpus selection and tokenization strategies significantly influence abusive language detection performance. Tamil-specific and Indic-language pretrained models were generally more robust toward noisy social media text, spelling variations, transliterated Tamil-English expressions, and contextual conversational meaning.

5.3 Discussion

Tamil-BERT achieves the highest accuracy of 80.72% largely because it was trained exclusively on Tamil text, giving it unmatched morphological precision for the language. This behavior is also reflected in the confusion matrix (Figure 2), which shows fewer misclassifications between abusive and non-abusive classes. MuRIL follows closely at 79.19%, drawing its strength from a corpus spanning 17 Indian languages with transliterated pairs, essentially giving it a natural feel for informal "Tanglish" on social media. Models like mBERT and XLM-RoBERTa, trained on 100+ language global corpora, offer a solid baseline but lack that regional depth. IndicBERT, though focused on 12 Indian languages, misses the transliteration-aware data that gives MuRIL its edge.

On tokenization, mBERT and MuRIL use WordPiece, which effectively separates roots from suffixes, while Tamil-BERT, XLM-RoBERTa, and IndicBERT use SentencePiece, better suited to the morphological complexity of agglutinative languages like Tamil. Among the models, Tamil-BERT achieves the highest accuracy of 0.8072, followed by MuRIL (0.7919) and mBERT (0.7711), suggesting that both tokenization strategy and training data significantly influence performance.

Architecturally, the BERT-based models, Tamil-BERT, mBERT, and MuRIL, benefited most from bidirectional encoding for this task. XLM-RoBERTa's RoBERTa backbone brings greater scale but not enough regional specificity to compete with the top three. IndicBERT's ALBERT architecture, while efficient and lightweight, likely fell short due to its reduced parameter capacity, limiting its ability to detect the more subtle patterns of abusive language.

6 Conclusion

This paper investigates Tamil abusive content classification as a binary sequence classification task. We fine-tuned and evaluated five transformer-based pretrained language models—mBERT, MuRIL, XLM-RoBERTa, IndicBERT, and Tamil-BERT. Our results indicate that language models pretrained on Indic or Tamil-specific corpora demonstrate better linguistic alignment compared to generic multilingual language models. Among the evaluated models, Tamil-BERT achieved the best performance with an accuracy of 80.72%, highlighting the advantage of language-specific pretraining for Tamil abusive content detection.

Future work can focus on further improving Tamil-BERT through techniques such as domain-adaptive pretraining on larger Tamil social media corpora and incorporating data augmentation strategies for abusive language variations. Additionally, integrating contextual features such as code-mixed Tamil-English patterns and leveraging ensemble or hybrid architectures could further enhance model robustness and classification accuracy.

7 Limitations

While the proposed transformer-based approaches achieved promising performance on Tamil abusive content detection, there are a few challenges that remain to be addressed. First, despite utilizing state-of-the-art pretrained multilingual models, the size of the dataset is a few order of magnitude smaller than those used for multilingual NLP tasks. This can potentially impede the models' ability to generalize to unseen data. Furthermore, social media text contains a lot of noisy, often sarcastic, and context-dependent language that is very challenging to classify accurately. There is also a continued threat of new slang and abusive language that has not been seen and annotated in the training data.

While multilingual transformer models are powerful for learning contextual understanding that generalizes across languages, dialectalvariations within a language such as Tamil spoken in different geographical distribution of Tamil-speaking users are challenging to address. We also note significant increases in computational complexity and inference costs with scaling up the transformer model size for real-time moderation use cases.

8 Ethical Considerations

This paper aims to detect abusive content in order to improve online safety and reduce harm caused by online interactions. Developing a system for abusive content detection requires abusive language datasets, which can be full of offensively written content. The data used here for training and testing a learning-based approach were taken from a publicly available shared task resource, and used purely for research purposes.

While automated systems for detecting abusive language can miss some instances of abuse or flag some non-abusive language (false positives and false negatives, respectively)—especially when the language includes sarcasm, makes reference to the culture of a user or group not notified in the system's training data, or is worded implicitly rather than directly—such automated systems are best used as a helper for human moderators, highlighting potentially abusive language for them to decide upon. Where possible, the models should also be evaluated for bias, to ensure that the system does not unfairly flag users speaking from certain linguistic, cultural, dialectical, or stylistic backgrounds.

9 AI Usage Disclosure

AI tools were used during the preparation of this manuscript to assist with language editing, formatting of LaTeX content, and improving clarity of presentation. The core research contributions, experimental design, model implementation, and result analysis were conducted entirely by the authors. All generated suggestions were carefully reviewed and verified by the authors before inclusion in the final manuscript.

References

A. Abishek and T. C. Govindapillai. 2024. Detecting tamil textual threats in social media using artificial intelligence. *International Journal of Novel Research and Development (IJNRD)*, 9(3):26–33.

M. Alam and S. Rahman. 2025. Cuet_ignite@dravidianlangtech 2025: Detection of abusive comments in tamil text using transformer models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 411–418. Association for Computational Linguistics.

S. Bhattacharyya, P. Gupta, and M. Lowe. 2025. Lexilogic@dravidianlangtech 2025: Detecting misogynistic memes and abusive tamil and malayalam text targeting women on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 112–119. Association for Computational Linguistics.

S. U. Hegde and A. Hande. 2025. Team_strikers@dravidianlangtech 2025: Misogyny meme detection in tamil using multimodal deep learning. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 605–612. Association for Computational Linguistics.

A. Kalaivani and D. Thenmozhi. 2024. Composite feature fusion for improved offensive language detection in tamil social media using mha-lstm. *International Journal of Machine Learning and Cybernetics*, 15:1–15.

P. Kumaresan and S. Karnati. 2025. Kec_ai_vss_run2@dravidianlangtech 2025: Abusive tamil and malayalam text targeting women on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 315–322. Association for Computational Linguistics.

A. Mallik, R. Dhar, U. Das, M. A. Labib, S. Rahman, and H. Murad. 2025. Cuet-823@dravidianlangtech 2025: Shared task on multimodal misogyny meme detection in tamil language. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language*

Technologies for Dravidian Languages, pages 401–408. Association for Computational Linguistics.

J. Mohan, S. R. Mekapati, B. Premjith, G. J. Lal, and B. R. Chakravarthi. 2025. 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*, 24(3):1–24.

J. Pokrywka and K. Jassem. 2024. Evaluating transformer models for hate speech detection in tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 196–199. Association for Computational Linguistics.

R. Ponnusamy, B. Sivagnanam, 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)*, pages 660–671. ELRA and ICCL.

S. Rajiakodi, B. R. Chakravarthi, S. P. M. Chinnan, R. Priyadharshini, J. Rajameenakshi, and 1 others. 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*, pages 671–681. Association for Computational Linguistics.

S. Sangeetham and M. Bedi. 2025. Jas@dravidianlangtech 2025: Abusive tamil text targeting women on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 210–216. Association for Computational Linguistics.

M. Subramanian, K. S. Vadivel, and R. Sowmya. 2022. Detecting offensive tamil texts using machine learning and multilingual transformer models. In *Proceedings of the 2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, pages 1–6. IEEE.

U. Thayasivam and Y. S. Wijesuriya. 2025. Gs_dravidianlangtech@2025: Women targeted abusive texts detection on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 520–527. Association for Computational Linguistics.