DLRG@DravidianLangTech 2026: Explainable Transformer-Based Detection of Abusive Tamil Text Targeting Women on Social Media

Mirudhula Sankar and Ratnavel Rajalakshmi
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, India
Corresponding Author: rajalakshmi.r@vit.ac.in

Abstract

Many social media platforms have users who have normalized the abuse of women online, creating a need for systems that automatically detect such activity. For low-resource, regional languages like Tamil, which has informal writing styles, spelling variations, dialectal differences, and culturally specific expressions, it becomes a challenge to correctly detect abusive comments. In this work, a transformer-based approach for binary classification of Tamil comments into abusive and non-abusive categories is done using the DravidianLangTech dataset. The proposed system fine-tunes MuRIL(a multilingual transformer pretrained for Indian languages), enabling effective contextual representation with minimal preprocessing. To improve the transparency of the system, a post-hoc Explainable AI component is incorporated. A perturbation-based method using log-odds differences identifies words that significantly influence the predictions. Experimental findings indicate that the model reaches a validation accuracy exceeding 81% while also exhibiting a strong macro-F1 score. This research shows that utilizing contextual multilingual representations alongside simple interpretability methods offers a viable and effective approach for detecting abusive text in Tamil. The implementation of our system is publicly available at https://github.com/mirud5173/Abusive-Tamil-Comment-Detection-using-Transformer-Models

1 Introduction

Social media platforms play a crucial role in communication and public engagement. However, they also contain a substantial amount of abusive and harmful material. Women face heightened risks of harassment, inappropriate language, and derogatory remarks. This exposure can cause psychological harm and decrease their involvement in online environments. Consequently, automated detection systems are vital for effective and scalable content moderation.

Research on detecting abusive language has been extensively conducted in English and other well-resourced languages. However, Tamil presents additional challenges due to its low-resource nature. Social media text in Tamil is highly informal and includes phonetic spellings, slang, dialectal variations, and inconsistent grammar. Furthermore, abusive intent may be expressed implicitly through sarcasm, cultural references, or metaphorical expressions.

Transformer-based models have significantly improved performance in text classification tasks. Multilingual pretrained models are particularly useful for low-resource settings as they capture cross-lingual semantic representations. The DravidianLangTech shared tasks provide benchmark datasets that enable research on abusive language detection for Tamil.

This work proposes a MuRIL-based system for detecting abusive Tamil text targeting women. In addition to strong classification performance, a perturbation-based explainability module is introduced to improve interpretability and trust in model decisions.

The dataset and evaluation protocol follow the shared task findings report (Rajiakodi et al., 2026).

2 Related Work

The task of detecting abusive Tamil text on social media has gained significant attention through the

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 237–241

DravidianLangTech shared tasks, which provide benchmark datasets and evaluation frameworks for abusive and offensive language detection in Dravidian languages (Rajiakodi et al., 2026). Earlier work in this domain relied on traditional machine learning approaches using handcrafted features such as n-grams, TF-IDF vectors, and lexical cues. While these methods proved to be computationally efficient, they faced challenges in recognizing contextual subtleness and underlying abusive intent in casual social media language.

With the rise of deep learning techniques, more recent research has increasingly favored transformer-based architectures, which excel at capturing contextual semantic nuances. Models like mBERT and XLM-RoBERTa have been prominently used in detecting abusive language in Tamil and other Dravidian languages, also for the task of sentiment analysis (Naib et al., 2025; Rajiakodi et al., 2025; Hanif and Rahman, 2025; Kannan et al., 2021; Rajalakshmi et al., 2022b; Sivakumar and Rajalakshmi, 2021). These models have notably enhanced detection performance by adeptly addressing informal language, spelling variations, and code-mixed text typically found on social media.

Previous research (Rajalakshmi et al., 2022a) examined multilingual transformer models for identifying abusive comments in Tamil, showing that pretrained multilingual models can effectively discern linguistic patterns in languages with limited resources. Similarly, another study (Rajalakshmi et al., 2021) focused on offensive language detection in code-mixed Tamil through transformer-based methods, emphasizing the significance of contextual embeddings in managing multilingual and code-mixed content. These investigations provided initial support for the idea that transformer-based architectures surpass traditional machine learning approaches in detecting abusive language in Tamil.

To boost classification performance, hybrid architectures that merge transformer embeddings with traditional machine learning classifiers have also been considered. For example, systems that integrate RoBERTa with ensemble classifiers like XGBoost have shown improved generalization by utilizing both deep contextual representations and supplementary classification layers (Nishanth et al., 2025).

In addition to large transformer models, lightweight architectures such as FastText and shallow neural networks have been investigated to reduce computational complexity and enable faster inference. While these approaches provide efficiency benefits, they generally achieve lower performance compared to contextual transformer-based systems, emphasizing the importance of deep semantic modeling for abusive language detection (Rohit et al., 2025).

Several studies have also highlighted the role of preprocessing strategies, hyperparameter tuning, and ensemble learning in improving system robustness and classification performance (Radha and Swathika, 2025; Subramanian and Shanmugavadivel, 2025). Although ensemble-based systems can provide marginal improvements in benchmark scores, they often increase model complexity and deployment overhead.

Overall, existing work primarily focuses on maximizing classification performance within the shared task benchmark, where macro-F1 scores typically range between 0.70 and 0.82. However, many of these systems function as black-box models and offer limited interpretability regarding their predictions.

In contrast, the proposed work aims to achieve competitive performance using a single transformer-based architecture while integrating a perturbation-based explainable AI component. This approach enables identification of influential words contributing to model predictions, improving transparency and making the system more suitable for real-world content moderation applications.

3 Proposed Method

3.1 System Overview

The comprehensive architecture of the proposed system is shown in Figure 1. The architectural pipeline involves three phases: basic preprocessing, transformer-based classification, and post-hoc explanation.

3.2 Data Preparation

The dataset features Tamil social media comments classified as either abusive or non-abusive, with only simple preprocessing applied. The text remains in its original state, avoiding extensive normalization or the removal of stopwords. This approach enables the model to capture contextual and stylistic nuances from authentic user-generated content.

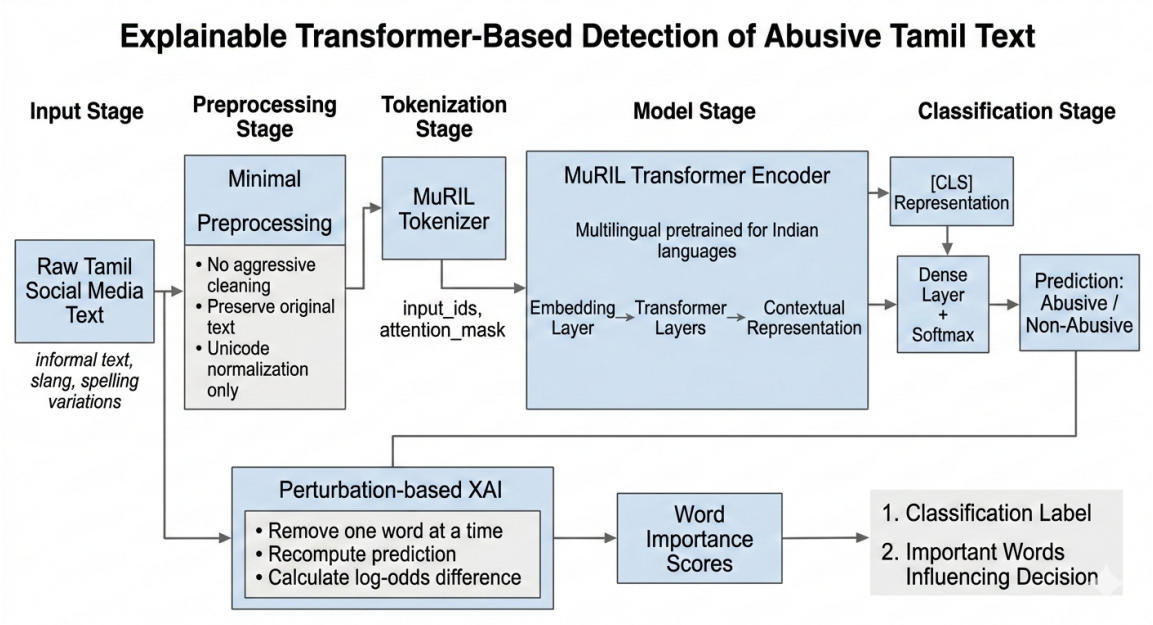

Figure 1: Overall framework of the proposed system. Tamil text is minimally preprocessed and tokenized using MuRIL. The transformer generates contextual representations used for binary classification. A perturbation-based module computes word importance scores for explanation.

3.3 Model Training

MuRIL serves as the foundational encoder, where the input text is tokenized and transformed into contextual embeddings. The representation of the [CLS] token is processed through a fully connected layer for binary classification. The model undergoes fine-tuning using cross-entropy loss, set with a learning rate of 2×10^{-5} and a batch size of 16, while validation metrics are monitored to identify the optimal model.

3.4 Explainable AI Component

For each input sentence, the initial prediction probability is calculated. Subsequently, individual words are removed one at a time, and the prediction is reassessed. The importance score for each word is defined as follows:

$$Importance_i = \log \frac{p}{1-p} - \log \frac{p_i}{1-p_i}$$

Words with higher importance scores have a greater influence on the final decision. This approach provides intuitive explanations without modifying the model architecture.

4 Experimental Results and Discussion

4.1 Training Performance

Table 1 shows steady improvement in both accuracy and macro-F1 during early epochs, followed by stabilization after convergence.

Epoch	Accuracy	Macro F1
1	0.7141	0.7135
2	0.7688	0.7687
3	0.7784	0.7777
4	0.7907	0.7904
5	0.7934	0.7923
6	0.8085	0.8069
7	0.8140	0.8126
8	0.8030	0.8026
9	0.8071	0.8070
10	0.8057	0.8049

Table 1: Validation performance across training epochs.

4.2 Classification Analysis

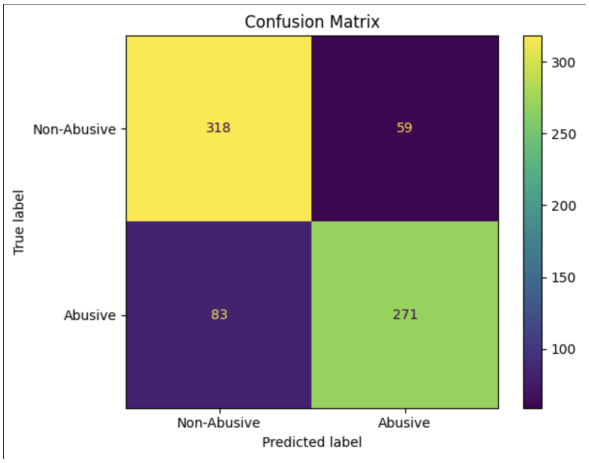

Figure 2: Confusion matrix of the proposed model.

Model	Accuracy	Macro F1
TF-IDF + SVM	0.7975	0.7975
MuRIL (baseline)	0.7800	0.7809
Hybrid (TF-IDF + MuRIL)	0.7900	0.7934
Proposed MuRIL + XAI	**0.8140**	**0.8126**

Table 2: Performance comparison of evaluated models.

The confusion matrix indicates balanced performance across both classes. Most errors occur in cases where abusive intent is implicit or context-dependent.

4.3 Explainability Analysis

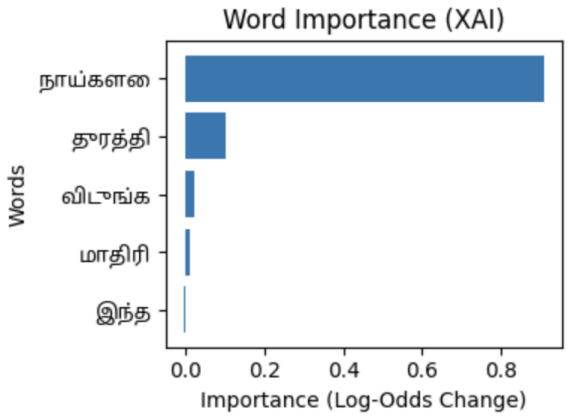

Figure 3: Example word-level importance generated by the explainability module.

The explanation results show that the model focuses on semantically meaningful abusive expressions, indicating effective contextual learning.

5 Comparative Study

Table 2 shows that the proposed system achieves the best performance among all evaluated approaches. The results demonstrate that contextual transformer representations combined with careful fine-tuning provide better generalization than traditional and hybrid methods. Additionally, the integration of explainability improves transparency without affecting performance.

6 Conclusion

This work, thus, presents an explainable transformer-based system for detecting abusive Tamil text targeting women. Fine-tuning MuRIL enabled effective contextual understanding of informal social media language while requiring minimal preprocessing. The proposed model achieved validation accuracy above 81% with strong macro-F1 performance.

A perturbation-based explainability method was introduced to identify influential words contributing to model decisions. This improves interpretability and supports responsible deployment in automated moderation systems.

Future work includes extending the approach to multi-class abusive categories, handling code-mixed Tamil-English content, and exploring advanced explanation techniques for deeper interpretability.

7 Limitations

While the proposed approach shows promising results, several limitations remain. The current model performs only binary classification and does not distinguish between different categories or severity levels of abusive language. In addition, the experiments rely on a shared-task dataset of Tamil social media comments, which may limit the model's generalization to other platforms or domains. The explanation module estimates word importance by removing one word at a time, which provides an intuitive interpretation but may not fully capture complex contextual relationships between words. Furthermore, the system depends on the MuRIL pretrained model, and its performance may vary across different datasets or low-resource settings. Future work can explore multi-class abusive categories, larger datasets, and more advanced explanation techniques.

8 Generative AI Acknowledgment

Generative AI tools were used to assist with portions of code generation and early-stage drafting. All generated content was carefully reviewed, validated, and substantially revised by the authors.

9 Ethical Considerations

This work focuses on detecting abusive Tamil text targeting women on social media. While automated systems can help reduce harmful content, machine learning models may still misclassify text and should support rather than replace human moderators.

The dataset consists of publicly available DravidianLangTech social media comments with no personally identifiable information analyzed. This work aims to support safer online spaces using transparent and explainable AI techniques.

References

G. Dhanyashree and 1 others. 2025. Linguaists@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*.

Tareque Md Hanif and Md Rashadur Rahman. 2025. Cuet_agile@dravidianlangtech 2025: Fine-tuning transformers for detecting abusive text targeting women from tamil and malayalam texts. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*.

R Ramesh Kannan, Ratnavel Rajalakshmi, and Lokesh Kumar. 2021. Indicbert based approach for sentiment analysis on code-mixed tamil tweets. In *FIRE (Working Notes)*, pages 729–736.

Md Mubasshir Naib and 1 others. 2025. cuetraptors@dravidianlangtech 2025: Transformer-based approaches for detecting abusive tamil text targeting women on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*.

S. Nishanth and 1 others. 2025. Ansr@dravidianlangtech 2025: Detection of abusive tamil and malayalam text targeting women on social media using roberta and xgboost. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*.

N. Radha and R. Swathika. 2025. Ssn_it_nlp@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*.

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022a. 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. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Preethi Reddy, Shreya Khare, and Vaishali Ganganwar. 2022b. Sentimental analysis of code-mixed hindi language. In *Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 2*, pages 739–751. Springer.

Ratnavel Rajalakshmi, Yashwant Reddy, and Lokesh Kumar. 2021. Dlrg@dravidianlangtech-eacl2021: Transformer based approach for offensive language identification on code-mixed tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 357–362. Association for Computational Linguistics.

Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Ratnavel Rajalakshmi, Kathiravan Pannerselvam, Bhuvaneswari Sivagnanam, V Jananayagan, 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.

Saranya Rajiakodi and 1 others. 2025. Findings of the shared task on 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*.

V. P. Rohit and 1 others. 2025. Cyber protectors@dravidianlangtech 2025: Abusive tamil and malayalam text targeting women on social media using fasttext. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*.

B. Saathvik, Janeshvar Sivakumar, and Thenmozhi Durairaj. 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*.

Kogilavani Shanmugavadivel and Malliga Subramanian. 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*.

Soubraylu Sivakumar and Ratnavel Rajalakshmi. 2021. Self-attention based sentiment analysis with effective embedding techniques. *International Journal of Computer Applications in Technology*, 65(1):65–77.

Malliga Subramanian and Kogilavani Shanmugavadivel. 2025. Kecempower@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*.