HNK@DravidianLangTech 2026: Investigating Grapheme-Level Normalization for Abusive Tamil Text Classification

Hanish Vigneshwar R, Nahul Alaguraj, Karthikeyan M, and Ratnavel Rajalakshmi
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, India
Corresponding Author: rajalakshmi.r@vit.ac.in

Abstract

The increasing prevalence of social media has also correlated with an increase in abusive content targeting women, particularly for regional languages such as Tamil. The automatic identification of abusive content is critical for the creation of safer online spaces. In this paper, we focus on the abusive text detection of women in the context of binary text classification. We evaluated the performance of the proposed system on the abusive text detection of women using the IndicBERT, MuRIL, and Tamil-BERT models. Additionally, we propose the use of grapheme-aware normalization for the proposed system. Grapheme-aware normalization aims to maintain the structural integrity of Tamil characters at the Unicode level. The experimental results reveal that the proposed system using the Tamil-BERT model with grapheme-aware normalization achieves the best performance among the evaluated models. The proposed system achieved the third position in the shared task. [1]

1 Introduction

The rapid growth of social media has transformed online communication by enabling large-scale interaction and information exchange. However, these platforms have also become spaces where abusive language is frequently directed towards women, leading to significant psychological and social consequences. This increasing prevalence of online abuse highlights the need for automated systems capable of identifying harmful content.

Although abusive language detection has been widely studied for high-resource languages such as English, research for low-resource languages like Tamil remains limited. Tamil is a morphologically rich Dravidian language with a complex writing system, agglutinative structure, and frequent code-mixing with English, which makes abusive language detection challenging. Recent studies have shown that transformer-based models can improve Tamil abusive text classification performance (Rajalakshmi et al., 2023; Hanif and Rahman, 2025).

Transformer-based models such as IndicBERT, MuRIL, and TamilBERT have advanced natural language processing for Indian languages by capturing contextual information from multilingual corpora. However, their performance is influenced by preprocessing and tokenization quality. Prior work highlights the importance of tokenization for representing morphologically rich and low-resource languages effectively (Velayuthan and Sarveswaran, 2023). Preserving consistent Tamil character representation at the Unicode and grapheme level is therefore important for improving model understanding.

In this work, abusive Tamil text targeting women is treated as a binary classification task. We evaluate IndicBERT, MuRIL, and TamilBERT under standard and grapheme-aware preprocessing settings. Unlike prior work that primarily focuses on tokenizer analysis, our approach investigates grapheme-aware normalization as a preprocessing strategy for Tamil abusive text classification. Experimental results show that grapheme-aware normalization is most beneficial for TamilBERT, and the proposed system secured third place in the shared task.

2 Related Work

Abusive language detection on social media has been widely studied, particularly for high-resource languages such as English. Early approaches relied on traditional machine learning models including Support Vector Machines (SVM) and Logistic Regression using handcrafted features such as n-grams and lexical patterns. While these methods

[1] Code and implementation are available at: https://github.com/The-Silly-Glitch/GraphemeAware-Tamil-Abuse-Detection

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 258–262

provided strong baselines, they struggled to capture contextual and implicit forms of abusive language (Mozafari et al., 2024).

Deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks later improved contextual modeling for abusive language detection. However, their effectiveness in low-resource and multilingual settings remained limited due to linguistic variability and insufficient annotated data (Mnassri et al., 2024).

Transformer-based models have since become the dominant approach for text classification tasks. Models such as BERT and its variants leverage self-attention mechanisms to capture contextual relationships effectively (Ashish, 2017). Multilingual transformer models including mBERT, XLM-RoBERTa, IndicBERT, and MuRIL have shown strong performance across several Indian language processing tasks.

Research on abusive language detection in Tamil has expanded through the DravidianLangTech shared tasks. Multilingual transformer models for abusive comment detection in Tamil was explored in (Rajalakshmi et al., 2022) and they demonstrated the effectiveness of fine-tuned transformer architectures. Earlier work (Rajalakshmi et al., 2021) investigated transformer-based methods for offensive language identification in code-mixed Tamil, highlighting the challenges of multilingual and informal social media text. Other studies have also examined transformer-based approaches and preprocessing strategies for Tamil abusive language detection (Hanif and Rahman, 2025).

Recent work has emphasized the importance of tokenization quality for morphologically rich and underrepresented languages. Velayuthan and Sarveswaran (Petrov et al., 2023) analyzed how tokenization strategies influence language representation in multilingual models. Unlike prior work that focuses primarily on tokenizer analysis and vocabulary representation, our work investigates grapheme-aware normalization as a preprocessing strategy for Tamil abusive text classification.

Although prior studies have explored transformer-based models for Tamil abusive language detection, comparatively fewer works examine grapheme-level normalization and character representation consistency for Tamil text.

3 Dataset Description

The dataset used in this study was released as part of the shared task on detecting abusive Tamil text targeting women on social media (Sivagnanam et al., 2026). It consists of Tamil social media comments annotated under a binary classification setting as either *Abusive* or *Non-Abusive*.

Class	Instances
Non-Abusive	1883
Abusive	1768
Total	3652

Table 1: Class distribution of the training dataset.

The training split contains 3,652 Tamil comments, with 1,883 labeled as Non-Abusive and 1,768 labeled as Abusive. A small number of samples contained inconsistent label formatting, which was normalized during preprocessing. Overall, the dataset is relatively balanced, supporting stable model training and evaluation.

The dataset mainly consists of user-generated comments written in Tamil script. The comments frequently contain informal language, slang, spelling variations, and code-mixing with English words and numerals. These characteristics reflect real-world social media communication but also increase the complexity of automatic abusive language detection.

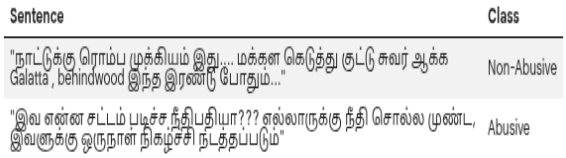

Figure 1: Example comments from the dataset illustrating non-abusive and abusive Tamil text.

Figure 1 presents example comments from the dataset. The examples illustrate the variation in writing style and the presence of implicit abusive expressions commonly found in social media text.

Class	Instances
Non-Abusive	472
Abusive	441
Total	913

Table 2: Class distribution of the test dataset.

For model development, the training data was further divided into training and validation subsets. The labeled test dataset contains 913 comments,

with 472 labeled as Non-Abusive and 441 labeled as Abusive, and was used for final evaluation.

4 Methodology

Abusive Tamil text detection is formulated as a supervised binary classification task, where each social media comment is classified as either *Abusive* or *Non-Abusive*. The overall workflow of the proposed system is shown in Figure 2.

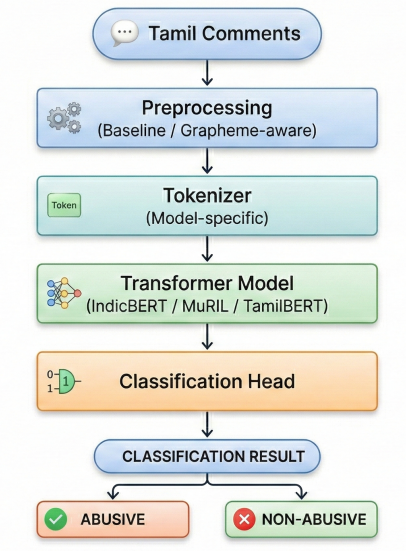

Figure 2: System workflow. Tamil text undergoes either baseline preprocessing or grapheme-aware normalization before classification using transformer-based models.

4.1 Preprocessing

Basic preprocessing steps including URL removal, whitespace normalization, and Unicode normalization were applied to ensure consistent Tamil text encoding. Since social media comments often contain irregular formatting and inconsistent character representations, normalization was necessary before training.

In addition to baseline preprocessing, we investigate a grapheme-aware normalization strategy for Tamil text. Tamil characters may consist of multiple Unicode code points that together form a single visible grapheme. Direct Unicode-level tokenization can therefore lead to inconsistent character segmentation. To address this issue, grapheme segmentation was applied before tokenization to preserve linguistically meaningful character boundaries.

Figure 3 compares TamilBERT tokenization with and without grapheme-aware preprocessing. Standard tokenization follows sub word segmentation directly on the input text, while grapheme-aware preprocessing preserves character boundaries before tokenization.

4.2 Models

Three transformer-based models were fine-tuned for the classification task: IndicBERT, MuRIL, and TamilBERT. IndicBERT and MuRIL are multilingual models pretrained on Indian languages, while TamilBERT is specifically pretrained on Tamil text. In all experiments, the original pretrained architectures were retained and only the classification layer was fine-tuned.

4.3 Training Setup

All models were trained using identical hyper parameters for fair comparison. A learning rate of 2×10^{-5}, batch size of 16, and 5 training epochs were used. The AdamW optimizer and cross-entropy loss were employed for training. The dataset was divided into training and validation splits. No data augmentation, class weighting, or ensemble methods were used.

4.4 Evaluation

Model performance was evaluated using the F1-score. Each model was evaluated under both preprocessing configurations, resulting in six experimental settings. The configuration achieving the highest validation F1-score was selected for generating predictions on the test dataset.

5 Results and Discussion

We evaluated IndicBERT, MuRIL, and TamilBERT under two preprocessing settings: baseline normalization and grapheme-aware normalization. Performance was first measured using the F1-score on the validation set.

Model	Baseline F1	Grapheme F1
IndicBERT	0.8201	0.8156
MuRIL	0.7958	0.7956
TamilBERT	0.7941	**0.8476**

Table 3: Validation F1-scores under baseline and grapheme-aware preprocessing.

As shown in Table 3, grapheme-aware normalization substantially improved TamilBERT, resulting in the highest overall validation F1-score. In contrast, IndicBERT and MuRIL showed little improvement under grapheme-aware preprocessing.

To further analyze model behavior, predictions from each model were evaluated on the labeled test dataset.

Representation	Output
Original Text	இவ என்ன சட்டம் படிச்ச நீதிபதியா
Tokenization (Without Grapheme)	['இ', '##வ', 'என்ன', 'சட்டம்', 'படி', '##ச்ச', 'நீதிபதி', '##யா']
Tokenization (With Grapheme)	['இ', 'வ', 'எ', 'ன்', 'ன', 'ச', 'ட', '##ம்', 'ட', 'ம்', 'ப', 'டி', 'ச', '##்', 'ச', 'நீ', 'தி', 'ப', 'தி', 'யா']

Figure 3: Comparison of TamilBERT tokenization with and without grapheme-aware preprocessing for a Tamil sentence.

Model	Base Acc.	Grapheme Acc.
IndicBERT	0.7919	0.8039
MuRIL	0.7766	0.7777
TamilBERT	0.8018	**0.8105**

Table 4: Test set accuracy under baseline and grapheme-aware preprocessing.

The test results follow a similar trend. Grapheme-aware normalization improved TamilBERT and slightly improved IndicBERT, while MuRIL showed minimal change. TamilBERT with grapheme-aware preprocessing achieved the best overall test accuracy.

One possible explanation is the difference in pretraining objectives and vocabulary construction across the models. TamilBERT is pretrained specifically on Tamil text and therefore benefits more from consistent grapheme-level character representation. In contrast, IndicBERT and MuRIL are multilingual models trained across multiple Indian languages and scripts, which may already provide greater robustness to Unicode and tokenization variations.

5.1 Error Analysis

A qualitative analysis of misclassified samples revealed several common error patterns. Implicit or context-dependent abusive comments were often misclassified as non-abusive, especially when explicit offensive terms were absent. Code-mixed Tamil-English comments also introduced tokenization inconsistencies. These observations suggest that although grapheme-aware normalization improves character representation, implicit meaning and contextual understanding remain challenging for automated systems.

Based on the validation results, TamilBERT with grapheme-aware preprocessing was selected for the shared task test set, where the proposed system secured third place.

6 Conclusion

This work explored abusive Tamil text detection targeting women on social media using transformer-based models. IndicBERT, MuRIL, and TamilBERT were evaluated under both standard and grapheme-aware preprocessing settings. The results show that multilingual models such as IndicBERT and MuRIL remain relatively stable across preprocessing methods, while TamilBERT benefits more from grapheme-aware normalization.

TamilBERT with grapheme-aware preprocessing achieved the best overall performance, obtaining a test accuracy of 0.8105 and securing third place in the shared task. Unlike prior work that primarily focuses on tokenizer analysis or multilingual representation learning, this study investigates grapheme-aware normalization as a preprocessing strategy for Tamil abusive text classification. The findings suggest that preserving grapheme-level character structure can improve representation quality for Tamil-specific transformer models.

Overall, the study highlights the importance of linguistically informed preprocessing for morphologically rich languages such as Tamil. Future work may explore larger datasets, additional transformer architectures, and more fine-grained abusive language categories.

Limitations

Although the proposed approach achieved promising results, several limitations remain. First, the experiments were conducted on a relatively limited shared-task dataset, which may affect generalization to broader social media contexts. Second, the study focuses on binary classification and does not capture finer-grained forms of abusive language such as sarcasm, implicit abuse, or harassment. Third, only three transformer-based models and

two preprocessing strategies were evaluated. Finally, abusive language is often context-dependent and culturally nuanced, which remains challenging for automated systems to model effectively.

References

Vaswani Ashish. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:I.

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*, pages 315–319, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Khouloud Mnassri, Reza Farahbakhsh, Razieh Chalehchaleh, Praboda Rajapaksha, Amir Reza Jafari, Guanlin Li, and Noel Crespi. 2024. A survey on multi-lingual offensive language detection. *PeerJ Computer Science*, 10:e1934.

Marzieh Mozafari, Khouloud Mnassri, Reza Farahbakhsh, and Noel Crespi. 2024. Offensive language detection in low resource languages: A use case of persian language. *PLoS one*, 19(6):e0304166.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *ArXiv*, abs/2305.15425.

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.

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

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Faerie Mattins, Pavitra Vasudevan, and Anand Kumar. 2023. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech and Language*, 78:101464.

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.

Menan Velayuthan and Kengatharaiyer Sarveswaran. 2023. Egalitarian language representation in language models: It all begins with tokenizers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.