# SUPERNOVA@DravidianLangTech 2026: Transformer and Ensemble Approaches for Abusive Tamil Text Detection Targeting Women

**Kiruthika K[1], Roahiyaa T[1], Premjith B[1]**
[1]Amrita School of Artificial Intelligence,
Amrita Vishwa Vidyapeetham, Coimbatore, India
cb.sc.u4aie24329@cb.students.amrita.edu
cb.sc.u4aie24043@cb.students.amrita.edu
b_premjith@cb.amrita.edu

## Abstract

Abusive language targeting women on Tamil social media is a growing concern that necessitates automated detection systems capable of handling low-resource, code-mixed, and morphologically rich text. This paper presents the SUPERNOVA system submitted to the shared task on Abusive Tamil Text Targeting Women on Social Media at DravidianLangTech@ACL 2026. We investigate three complementary approaches: (1) fine-tuning MuRIL with class balancing and label smoothing, (2) MuRIL contextual embeddings combined with XGBoost and decision threshold tuning, and (3) a lightweight ensemble of character-level TF-IDF and SentenceBERT features with Random Forest and Extra Trees. Our best system achieves an accuracy of 0.8007 and a macro F1-score of 0.7994, ranking 11th among all participating teams. These results highlight the effectiveness of multilingual transformer representations combined with ensemble techniques for the detection of abusive text on Tamil social networks. The code is publicly available at https://github.com/Kiruthi001/SuperNova-DravidianLangTech-ACL2026.

## 1 Introduction

The rapid growth of social media has created new avenues for hate speech and abusive language, particularly targeting women. Tamil, spoken by over 80 million people across India and Sri Lanka, is widely used on platforms such as YouTube, Twitter/X, and Facebook. While this digital expansion has enabled greater communication and visibility, it has also amplified instances of targeted harassment and misogynistic content directed at Tamil-speaking women.

Automatically detecting such abusive content is both essential and challenging. Tamil social media text is linguistically complex due to its agglutinative morphology, frequent code-mixing with English, and the use of both native Tamil script and Roman transliteration. Additionally, the informal and rapidly evolving nature of online language further complicates automatic moderation systems.

The DravidianLangTech@ACL 2026 shared task (Rajiakodi et al., 2026) addresses this issue by providing annotated Tamil social media data for binary classification (*Abusive* vs. *Non-Abusive*) specifically targeting women. To tackle this challenge, we investigate three complementary systems: (1) an end-to-end fine-tuned MuRIL model, (2) a hybrid MuRIL and XGBoost approach, and (3) a CPU-efficient ensemble integrating character-level TF-IDF and SentenceBERT features with classical classifiers. Our best system ranks 11th among participating teams, achieving an accuracy of 0.8007 and a macro F1-score of 0.7994, suggesting that Indic-pretrained transformers combined with ensemble strategies are effective for abusive Tamil text detection in social media settings.

## 2 Related Work

Abusive language detection in Dravidian languages has been explored across multiple DravidianLangTech shared tasks (Chakravarthi et al., 2020). Early systems relied on classical methods such as TF-IDF with SVM and Random Forests, which showed competitive performance for morphologically rich languages. However, the emergence of pre-trained multilingual transformers significantly advanced the state of the art. mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) demonstrated strong cross-lingual transfer, while MuRIL (Khanuja et al., 2021), trained on 17 Indian languages, consistently outperformed generic models on Indic benchmarks. Furthermore, SentenceBERT (Reimers and Gurevych, 2019) provided efficient semantic embeddings that complement ensemble classifiers in resource-constrained settings. Our work builds on these findings by combining transformer-based representations with ensemble



*Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 376–380

strategies tailored for Tamil abusive text detection.

## 3 Methodology

Our methodological design aims to compare contextual deep learning approaches with classical machine learning models under a unified preprocessing and evaluation framework for abusive Tamil text detection.

### 3.1 Dataset

The dataset was released as part of the shared task on *Abusive Tamil Text Targeting Women on Social Media* at DravidianLangTech@ACL 2026 (Rajiakodi et al., 2026). It consists of Tamil YouTube comments annotated as *Abusive* or *Non-Abusive* by domain experts following predefined annotation guidelines. The dataset contains native Tamil script, Roman transliterations, and code-mixed Tamil-English content, reflecting the diversity of real-world social media language. Table 1 summarizes the dataset statistics.

| Split | Total | Abusive | Non-Abusive |
|---|---|---|---|
| Train | 25945 | 12971 | 12974 |
| Test | 913 | 441 | 472 |

Table 1: Dataset statistics.

### 3.2 System Architecture

Figure 1 illustrates the overall pipeline shared across all three runs. The raw Tamil text first passes through a common preprocessing stage. The cleaned text is then fed into three independent modeling pipelines, each representing a distinct approach. All three systems are evaluated using the same classification metrics defined by the shared task organizers. Character-level TF-IDF is particularly useful for Tamil due to its rich morphology and frequent spelling variations in informal social media text.

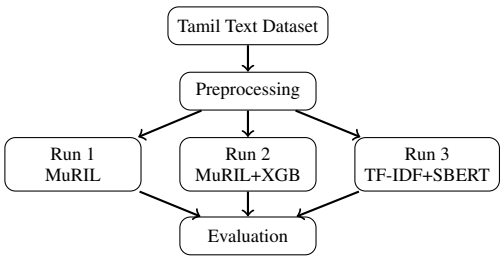


Figure 1: Overall system architecture.

### 3.3 Data Preprocessing

All input text undergoes the following steps: (1) lowercase normalization; (2) removal of URLs, @mentions, and #hashtags; (3) retention of Tamil Unicode characters (U+0B80–U+0BFF) and alphanumeric characters; and (4) collapsing of extra whitespace.

For transformer-based models, text is tokenized using subword tokenization with padding or truncation to a fixed maximum sequence length. For classical machine learning models, cleaned text is transformed into numerical TF-IDF feature vectors.

### 3.4 Run 1: MuRIL Fine-Tuning

We fine-tune `google/muril-base-cased` (Khanuja et al., 2021) for binary sequence classification. MuRIL is pre-trained on 17 Indian languages including Tamil, making it well-suited for this task. To address class imbalance, the abusive class is oversampled by duplication.

Training uses an 85/15 stratified split with the AdamW optimizer ($lr = 3 \times 10^{-5}$, weight decay $= 0.01$), cosine learning rate scheduling, label smoothing of 0.1, batch size of 16, and 6 training epochs with FP16 mixed precision. The maximum token length is set to 128.

This end-to-end fine-tuning strategy enables the model to adapt contextual representations specifically to abusive Tamil language patterns and context-dependent expressions.

### 3.5 Run 2: MuRIL Embeddings + XGBoost

MuRIL is first fine-tuned for 3 epochs using AdamW ($lr = 2 \times 10^{-5}$, maximum length $= 192$). Sentence-level embeddings are then extracted by concatenating the normalized `[CLS]` token representation and the mean-pooled last hidden state, resulting in a 1536-dimensional feature vector.

These embeddings are used to train an XGBoost classifier (Chen and Guestrin, 2016) with 4000 estimators, maximum depth of 4, and learning rate of 0.03. The classification threshold is tuned on the validation set to maximize macro F1-score.

This approach separates representation learning from classification, allowing tree-based boosting methods to operate on strong contextual embeddings without full end-to-end adaptation.

### 3.6 Run 3: TF-IDF + SentenceBERT + Ensemble

This configuration is designed to operate without GPU resources. It combines two complementary

feature types:

(a) Character-level TF-IDF (n-grams 3–5, maximum 50k features, min_df=5), which captures morphological patterns and is robust to transliteration noise.

(b) Multilingual SentenceBERT embeddings (`paraphrase-multilingual-MiniLM-L12-v2`, 384-dimensional) (Reimers and Gurevych, 2019), which provide cross-lingual semantic representations.

Both feature matrices are horizontally concatenated and passed to a soft-voting ensemble of Random Forest and Extra Trees classifiers (400 estimators each, class weights {0:1.0, 1:1.5}).

This configuration prioritizes computational efficiency and deployability in low-resource settings while maintaining reasonable classification performance.

### 3.7 Evaluation Metrics

Model performance is evaluated using Accuracy, Precision, Recall, and Macro F1-score in accordance with the shared task guidelines.

Accuracy measures overall correctness but may be misleading in the presence of class imbalance. Therefore, Macro F1-score is considered the primary evaluation metric, as it computes the F1-score independently for each class and then averages them, ensuring equal importance to both the *Abusive* and *Non-Abusive* categories.

Precision reflects how many predicted abusive instances are actually abusive, while Recall indicates how many abusive instances were successfully detected. The F1-score balances both metrics and provides a more comprehensive evaluation of model performance.

## 4 Experiments

### 4.1 System Configuration

Run 1 and Run 2 were conducted on Google Colab using an NVIDIA T4 GPU. Run 3 was executed on a CPU-only machine. All experiments use Python 3.10 with PyTorch, HuggingFace Transformers, Scikit-learn, and XGBoost libraries. The code is publicly available at `https://github.com/Kiruthi001/SuperNova-DravidianLangTech-ACL2026`.

### 4.2 Results

Table 2 presents the validation performance of all three runs on the training data. Table 3 presents the test set results for all three runs. Our Run 1 submission was officially ranked **11th** on the shared task leaderboard.

| Run | System | Acc | P | R | F1 |
|---|---|---|---|---|---|
| Run 1 | MuRIL Fine-tune | 0.8526 | 0.8383 | 0.8309 | 0.8283 |
| Run 2 | MuRIL+XGBoost | 0.8222 | 0.82 | 0.82 | 0.8216 |
| Run 3 | TF-IDF+SBERT+ET | 0.8085 | 0.81 | 0.81 | 0.8100 |

Table 2: Validation results on training data.

| Run | System | Acc | P | R | F1 |
|---|---|---|---|---|---|
| Run 1 | MuRIL Fine-tune | 0.8007 | 0.8031 | 0.7989 | 0.7994 |
| Run 2 | MuRIL+XGBoost | 0.8007 | 0.8031 | 0.7989 | 0.7994 |
| Run 3 | TF-IDF+SBERT+ET | 0.7415 | 0.7447 | 0.7432 | 0.7413 |

Table 3: Test set results. Run 1 ranked 11th officially.

### 4.3 Analysis

The validation results indicate that the transformer-based Run 1 model achieves the highest performance among all three approaches. With a validation macro F1-score of 0.8283, MuRIL fine-tuning demonstrates strong capability in modeling contextual dependencies in Tamil social media text. The use of cosine learning rate scheduling and label smoothing contributes to improved generalization during training.

However, a noticeable performance drop is observed when evaluating on the official test set (from 0.8283 to 0.7994 macro F1). This gap suggests potential domain shift between the training and test distributions or mild overfitting during fine-tuning. Despite this, Run 1 remains the strongest submission and achieves 11th rank on the leaderboard.

Run 2 (MuRIL + XGBoost) achieves slightly lower validation performance compared to Run 1. Although it leverages the same contextual representations, separating feature extraction and classification introduces additional optimization constraints. Unlike end-to-end fine-tuning, XGBoost does not update the transformer parameters during inference, which may limit adaptation to subtle abusive patterns.

Interestingly, Run 2 achieves identical official test scores to Run 1 in the leaderboard results. This suggests that tree-based classification over strong contextual embeddings can approximate end-to-end performance when decision thresholds are carefully tuned.

Interestingly, Run 2 achieves identical official test scores to Run 1 in the leaderboard results.

While Run 1 performs end-to-end fine-tuning, Run 2 relies on fixed MuRIL embeddings followed by an XGBoost classifier. This similarity suggests that the pretrained MuRIL representations are sufficiently expressive, allowing tree-based models to learn effective decision boundaries without full model adaptation. Additionally, the relatively small size of the test set and threshold tuning may contribute to the similar observed performance.

Run 3 demonstrates that classical machine learning approaches remain viable, especially under computational constraints. Character-level TF-IDF effectively captures Tamil morphological structures and spelling variations, including transliterated forms. The addition of SentenceBERT embeddings improves semantic coverage. However, the absence of deep contextual modeling limits its ability to detect implicit abuse, sarcasm, or context- dependent misogyny, resulting in lower macro F1-score (0.7413).

Overall, the results confirm that Indic-pretrained transformer models provide strong inductive bias for Tamil abusive text detection. While ensemble-based classical approaches offer computational efficiency, end-to-end contextual fine-tuning yields the most robust and consistent performance.

## 5 Conclusion

This paper presented three complementary systems for abusive Tamil text detection targeting women in the DravidianLangTech@ACL 2026 shared task. We compared transformer-based fine-tuning, hybrid embedding-based gradient boosting, and classical ensemble approaches under a unified evaluation framework.

Our best-performing system, based on end-to-end MuRIL fine-tuning, achieved an accuracy of 0.8007 and a macro F1-score of 0.7994, securing 11th rank on the official leaderboard. The results demonstrate that Indic-pretrained transformer models provide strong contextual representations for modeling code-mixed and morphologically rich Tamil social media text.

Additionally, our CPU-efficient ensemble model highlights that competitive performance can be achieved without GPU resources, making it suitable for practical moderation systems in low-resource environments.

Future work will explore data augmentation, context-aware modeling, and larger Indic language models to further enhance abusive text detection performance.

## Limitations

Despite achieving competitive performance, our models have several limitations. First, Tamil social media language evolves rapidly, including the use of new slang, emojis, and creative spellings, which may not be adequately captured by models trained on static datasets. Second, transformer-based approaches require GPU resources for efficient training, which may limit reproducibility in low-resource environments. Additionally, we did not explore advanced data augmentation techniques such as back-translation or cross-lingual transfer learning, which could potentially improve generalization performance. Finally, our models operate at the comment level and do not incorporate conversation context, which may be necessary for detecting implicit or context-dependent abuse.

## Ethical considerations

This work focuses on detecting abusive language targeting women in Tamil social media. Automated moderation systems must be used carefully to avoid bias or incorrect labeling of benign content as abusive. The models developed in this study are intended to assist human moderators rather than replace them. We use only the anonymized dataset provided by the shared task organizers and emphasize responsible deployment that respects user privacy and fairness.

## References


Bharathi Raja Chakravarthi and 1 others. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of SLTU-CCURL 2020*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of KDD 2016*.

Alexis Conneau and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*.

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 2019*.

Simran Khanuja and 1 others. 2021. MuRIL: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP 2019*.