CUET_SYNTHETICA@DravidianLangTech 2026: Multi Architecture Transformer Ensemble for Detecting Abusive Tamil Text Targeting Women

Miftahul Jannat Rishta, Sumaiya Zaman, Shiti Chowdhury, Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2104019, u2104110, u2004027}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Abusive language targeting women has been a serious problem on Tamil social media and building systems to detect it automatically is harder than it looks. Tamil is morphologically complex, people have written it mixed with English in ways no dictionary has accounted for and a lot of the hostility has been indirect enough that has slipped past models trained on surface patterns. In the Shared Task on Abusive Tamil Text Targeting Women on Social Media DravidianLangTech@ACL 2026, we have worked on classifying Tamil YouTube comments as Abusive or Non-Abusive. We have trained three transformer models four times each with different learning rates, giving us 12 models total. Their predicted probabilities have been averaged to make the final decision. The 12-model ensemble has achieved a macro F1 of 0.8086, outperforming all individual models and securing 4th place in the shared task. Combining Tamil-specialized and multilingual transformer models has outperformed any single-architecture approach.

1 Introduction

Social media has made it far easier for people to connect but also far easier to harass them and women have borne more than their share of that harassment (Sivagnanam et al., 2026). Abusive comments have not just stay on screen rather they have pushed people out of conversations and have made online spaces feel hostile (Premjith et al., 2024). Detecting this content automatically is one concrete thing NLP can do about it.

Tamil has made that harder than it sounds. It is spoken by tens of millions yet has remained poorly covered in NLP research. The language is morphologically complex, users have written it heavily mixed with English and much of the abuse is culturally specific enough that surface-level models miss it entirely (Chakravarthi et al., 2022). Annotated Tamil datasets, which are also small in relation to the scale of the problem (Sreelakshmi et al., 2024). Despite these challenges, recent advancements in multilingual models and cross-lingual embeddings have shown promise in improving Tamil NLP but much work remains to be done to fully capture its complexities.

The DravidianLangTech@ACL 2026 shared task (Sivagnanam et al., 2026) has provided us a benchmark of Tamil YouTube comments to classify as Abusive or Non-Abusive. We have combined MuRIL (Khanuja et al., 2021) and IndicBERT v2 (Doddapaneni et al., 2023) for Tamil-specific coverage with XLM-RoBERTa-Large (Conneau et al., 2020) for a wider multilingual signal, having trained each architecture four times with different learning rates to get twelve models in total. Their predicted probabilities have been averaged for the final decision, achieving a macro F1 of 0.8086.

Our main contributions are:

- We have developed a heterogeneous ensemble of Tamil-specialized and multilingual transformers that has outperformed any single-model baseline.
- We have implemented learning-rate diversity across four runs per architecture to introduce prediction variance.

Code is available at: https://github.com/Mif-taa/Abusive-Tamil-Text.

2 Related Work

Most early work on abusive language detection has been done in English and the jump to Dravidian languages has been slow. Tamil social media text has proven genuinely difficult as users have mixed Tamil and English mid-sentence and have relied on cultural references that are hard to annotate systematically. Chakravarthi et al. (2022) has addressed this by releasing DravidianCodeMix, which has put Tamil, Malayalam and Kannada offensive language

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 212–216

detection on a shared benchmark for the first time. Dowlagar and Mamidi (2021) and Yasaswini et al. (2021) have shown that transformer fine-tuning can work for these languages, though class imbalance and code-mixing have remained difficult.

Indic-specific pretraining has made a real difference. Khanuja et al. (2021) has built MuRIL on 17 Indian languages and their transliterations - which matters because Tamil speakers on social media have often written Tamil words in Roman script. Doddapaneni et al. (2023) has extended this with IndicBERT v2 across 24 languages on a much larger corpus. Conneau et al. (2020) has shown that XLM-RoBERTa (Liu et al., 2019), scaled to 100 languages has transferred reasonably even to languages it has seen little of at training time. Sreelakshmi et al. (2024) has compared several of these models on Dravidian hate speech data and the DravidianLangTech shared tasks (Premjith et al., 2024; Sivagnanam et al., 2026) have pushed the field forward with standardized benchmarks. In the 2025 edition, Rahman et al. (2025) has found that combining MuRIL with XLM-BERT has outperformed either alone. Hanif and Rahman (2025) has found that learning rate choice matters more than model size and Thavarasa et al. (2025) has shown that implicit abuse in YouTube comments has been the hardest to catch - findings that have directly shaped our design. However, Tamil annotated data has remained limited and detecting abuse that relies on tone rather than explicit words has still been difficult, challenges we have faced in this work as well.

3 Data Description

The dataset has been sourced from the shared task on Abusive Tamil Text Targeting Women on Social Media-DravidianLangTech@ACL 2026. It has consisted of Tamil YouTube comments annotated with binary labels: Abusive and Non-Abusive. The training set has contained 3,652 samples and the test set 913. Some labels have had capitalization inconsistencies which we have corrected before training. The data distribution has shown in Table 1.

Label	Train	Test
Abusive	1,769	441
Non-Abusive	1,883	472
Total	**3,652**	**913**

Table 1: Dataset distribution after label normalization.

4 Methodology

4.1 Problem Formulation

Each Tamil YouTube comment has needed to be assigned one of two labels: Abusive or Non-Abusive. Rather than betting on a single model to make that call, we have trained $K = 12$ transformer classifiers and letting them vote. Each model i has produced a softmax probability $p_i(k \mid t)$ for comment t and class $k \in \{0, 1\}$. The final label is whichever class has received the highest average probability across all 12 models:

$$\hat{y} = \arg\max_k \frac{1}{K} \sum_{i=1}^{K} p_i(k \mid t) \quad (1)$$

This soft-voting strategy has kept individual model confidence in the picture rather than just counting votes, which has mattered when models have disagreed on borderline cases.

4.2 Preprocessing

The raw training data has contained 3,652 samples. After removing 401 duplicates, 3,251 clean samples have remained. An 80/20 stratified split has yielded 2,600 training and 651 validation samples. Deduplication has been applied to training data only. Each comment has been cleaned through an 11-step pipeline: HTML entity decoding, URL removal, email and mention removal, hashtag normalization, punctuation reduction, whitespace normalization, standalone number removal, quote normalization and zero-width character removal. Labels have been standardized to consistent casing. The class distribution has been near-balanced (51.3% Non-Abusive, 48.7% Abusive), so mild imbalance has been handled through class-weight adjustment rather than resampling.

4.3 Transformer Models

We have experimented with three transformer architectures. MuRIL has been pre-trained on 17 Indian languages and their transliterations, making it well-suited for Tamil social media text where users frequently romanize Tamil words. IndicBERT v2 has been trained on a large-scale Indic corpus covering 24 languages, providing strong monolingual Tamil representations. XLM-RoBERTa-Large has covered 100 languages and has contributed broad cross-lingual signal that complements the Indic-specialized models. Among individual runs, IndicBERT_2 has delivered the strongest single-model performance on the test set.

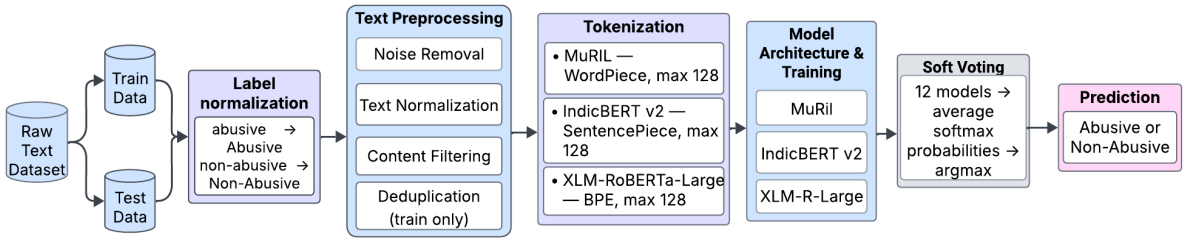

Figure 1: End-to-end pipeline of the heterogeneous 12-model ensemble for abusive Tamil text detection.

4.4 Ensemble

We have trained four runs per architecture by varying the learning rate, yielding 12 models total. Each model has been fine-tuned using an 80/20 train-validation split with class-weighted cross-entropy loss and the AdamW optimizer. Final predictions has been obtained by averaging the softmax probabilities across all 12 models. A soft-voting approach that accounts for prediction confidence rather than treating all votes equally. Figure 1 illustrates the overall process flow.

4.5 Evaluation Metrics

Macro F1, precision and recall have been used to evaluate all models, ensuring balanced performance across both the Abusive and Non-Abusive classes.

5 Results and Analysis

Table 2 has shown that among individual models, IndicBERT_2 has posted the strongest result F1: 0.8075, while the 12-model ensemble has outperformed all individual models with a macro F1 of 0.8086.

Model	Acc	P	R	F1
Best Single Models				
MuRIL_3	0.7963	0.7975	0.7949	0.7954
XLM-R-L_3	0.7919	0.7919	0.7913	0.7915
IndicBERT_2	0.8083	0.8096	0.8070	0.8075
Final Ensemble				
All 12 Models	**0.8094**	**0.8106**	**0.8082**	**0.8086**

Table 2: Best individual model per architecture and final 12-model ensemble on the official test set.

Table 3 has shown that no single-architecture ensemble has matched the full 12-model heterogeneous ensemble, confirming that architectural diversity has contributed beyond learning-rate variation alone. The final 12-model soft vote has pushed F1 to 0.8086, above any single model or single-architecture ensemble.

Architecture	Acc	P	R	F1
MuRIL (4 models)	0.7952	0.7949	0.7948	0.7948
XLM-R-L (4 models)	0.7996	0.7993	0.7991	0.7992
IndicBERT (4 models)	0.7985	0.8011	0.7964	0.7970
All 12 Models	**0.8094**	**0.8106**	**0.8082**	**0.8086**

Table 3: Ablation: per-architecture 4-model ensembles vs. final heterogeneous 12-model ensemble.

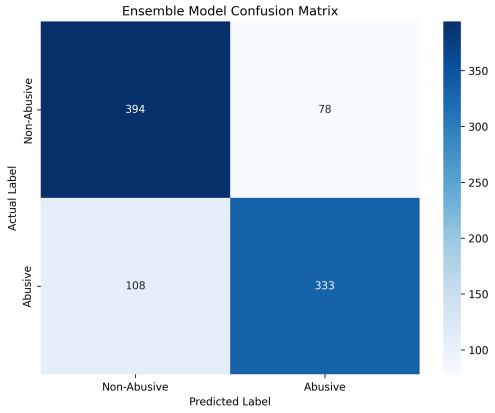

Figure 2: Confusion matrix for the final 12-model ensemble on the test set.

Figure 2 has shown the confusion matrix for the final ensemble. The model has correctly identified 394 non-abusive and 333 abusive comments, while producing 78 false positives and 108 false negatives. The higher false negative count suggests the ensemble has more often missed abusive content than over-flagged normal comments-a gap we would prioritize closing in future work.

5.1 Hyperparameter Settings

Table 4 has listed the training configuration for each model family. XLM-RoBERTa-Large has required much smaller learning rates (2e-6 to 5e-6) than MuRIL and IndicBERT v2. It is the largest model of the three and more sensitive to high learning rates. All models have used a cosine schedule with 10% warmup and AdamW with weight decay 0.01, training for 5/10 epochs.

Configurations	MuRIL	XLM-R-Large	IndicBERT v2
LR	4e-5	4e-6	3e-5
Batch	32	8	8
Epochs	10	10	5
Max Len	128	128	128
Optimizer	AdamW	AdamW	AdamW
W. Decay	0.01	0.01	0.01
Warmup	10%	10%	10%
Val Split	20%	20%	20%
Seeds	3407	3407	42

Table 4: Training configuration for 3 best model.

All experiments have been conducted on a Kaggle-provided NVIDIA Tesla T4 GPU (15.6 GB memory).

Architecture	R1	R2	R3	R4
MuRIL	2e-5	3e-5	4e-5	5e-5
XLM-R-Large	2e-6	3e-6	4e-6	5e-6
IndicBERT v2	2e-5	3e-5	2e-5	3e-5

Table 5: Learning rates across all 12 runs.

6 Conclusion

This work has addressed abusive Tamil text detection in the DravidianLangTech@ACL 2026 shared task, highlighting the challenges posed by morphological richness, code-mixing and culturally nuanced expressions. Our heterogeneous ensemble of MuRIL, XLM-RoBERTa-Large and IndicBERT v2 has outperformed all individual models on the official test set, demonstrating that combining Tamil-specialized and multilingual architectures has yielded more reliable predictions than relying on a single model. Although the performance gains have been consistent, they have been modest, indicating that learning-rate variation alone has provided limited ensemble diversity. Future research should explore greater model diversity through different random seeds, advanced ensemble strategies such as stacking or weighted voting, larger Tamil-focused pretrained models and improved modeling of implicit or culturally contextual abuse.

Error Analysis

The ensemble has struggled most with indirect and sarcastic abuse. The confusion matrix has shown 108 false negatives against 78 false positives, meaning the model has been more likely to miss abuse than over-flag normal content. False negatives have fallen into three categories: implicit abuse (hostile intent carried through cultural context), sarcastic comments (irony masking abusive intent) and code-mixed abuse (hostility appearing in the English portion of mixed comments). Representative false negatives have included comments containing culturally derogatory terms misread as neutral and sarcastic comments where tone has carried the hostility rather than explicit words. Future work should address these through sarcasm-aware modeling and larger Tamil datasets.

Limitations

The training set has been small at 3,251 samples with no labeled development set, so ensemble weights have been equal rather than tuned. Training 12 transformers is computationally expensive and all three architectures have shared similar pretraining objectives which has limited diversity and has likely kept gains modest. Full hyperparameter optimization using tools such as Weights & Biases has not been pursued due to the computational cost of running sweeps across 12 transformer models within the shared task timeline.

Ethical Statement

This work has used publicly available data from the shared task organizers. No personal information has been collected or stored. The dataset has contained harmful content targeting women and we have handled it carefully. We are aware that misclassification carries real costs and our goal has been fairness across both classes rather than metric optimization.

Acknowledgement

We have been grateful to the DravidianLangTech@ACL 2026 organizers for having provided the dataset and evaluation framework. We have thanked the developers of MuRIL, IndicBERT v2 and XLM-RoBERTa-Large whose open-source models have made this work possible. We have also acknowledged Devlin et al., Liu et al. and Chakravarthi et al. for the foundational work this research has built on.

References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2022. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 12402–12426. Association for Computational Linguistics.

Suman Dowlagar and Radhika Mamidi. 2021. OFFLangOne@DravidianLangTech-EACL2021: Transformers with the class balanced loss for offensive language identification in Dravidian code-mixed text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 154–159, Kyiv. Association for Computational Linguistics.

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*. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

Md Mizanur Rahman, Srijita Dhar, Md Mehedi Hasan, and Hasan Murad. 2025. MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL based transformer models for detection of 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*. Association for Computational Linguistics.

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.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Luxshan Thavarasa, Sivasuthan Sukumar, and Jubeerathan Thevakumar. 2025. Incepto@DravidianLangTech 2025: Detecting abusive Tamil and Malayalam text targeting women on YouTube. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194, Kyiv. Association for Computational Linguistics.