AbuseDetect_Alchemists@DravidianLangTech 2026: A Weighted Transformer Ensemble for Detecting Abusive Tamil Text Targeting Women

Meclin A Francis[1] **Jyoti Kumari**[2] **Vinay Babu Ulli**[3]
Malavika Sreekumar[4] Joel Johnson[5]
[1]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India
[2]Department of Linguistics, Banaras Hindu University, Varanasi, India
[3]Oogwai Analytics, Bangalore, India
[4]TransUnion, Pune, India [5]IBM, Kochi, India
meclinafrancis@gmail.com

Abstract

This paper describes our system submitted to the shared task on *Abusive Tamil Text Targeting Women on Social Media* at DravidianLangTech@ACL 2026. We formulate the problem as a supervised binary classification task, assigning each Tamil social media comment to an `Abusive` or `Non-Abusive` category. Our pipeline begins with a tailored preprocessing stage that handles emoji translation, URL removal, and entity normalization. We then independently fine-tune two pre-trained transformer models MuRIL and XLM-RoBERTa on the task data. At inference time, we combine these models through a weighted softmax ensemble, assigning a weight of 0.6 to MuRIL and 0.4 to XLM-RoBERTa. The resulting system achieves a Macro-F1 score of 0.8115 on the test set, outperforming both individual models. The code is publicly available.[1]

1 Introduction

Social media platforms have contributed to a significant increase in cyberbullying, hate speech, and targeted harassment (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Women face disproportionate levels of gender-based abuse intended to demean, intimidate, or suppress their online participation. In regional-language ecosystems such as Tamil, this abuse takes culturally specific forms, derogatory slang, sarcastic rebukes, and coded misogynistic remarks, that generic moderation tools fail to detect (Chakravarthi et al., 2021).

The DravidianLangTech@ACL 2026 shared task (Rajiakodi et al., 2026) on *Abusive Tamil Text Targeting Women on Social Media* requires systems to classify Tamil social media comments as `Abusive` or `Non-Abusive`. Tamil's agglutinative grammar, frequent code-mixing with English (*Tanglish*), and use of transliterated scripts introduce challenges that standard multilingual pretraining does not fully address.

We describe a classification pipeline with three key components: (1) a domain-specific preprocessing routine that normalizes noisy social media text while preserving emojis; (2) independent fine-tuning of MuRIL (Khanuja et al., 2021) and XLM-RoBERTa (Conneau et al., 2020); and (3) a weighted softmax ensemble that assigns higher confidence to MuRIL's stronger performance on Indic-language text while retaining XLM-RoBERTa's cross-lingual generalization.

2 Related Work

Early abuse detection relied on lexicons and classical classifiers (Schmidt and Wiegand, 2017), while pre-trained transformers such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) later advanced the state of the art. The DravidianLangTech workshop series has been instrumental in extending this research to Dravidian languages. Rajiakodi et al. (2025) provided the shared task overview, while Rahman et al. (2025) showed that MuRIL and XLM-BERT handle Tamil morpho-syntactic complexity effectively. Hanif and Rahman (2025) fine-tuned multilingual transformers; Kodali et al. (2025) combined XLM-RoBERTa with attention-driven BiLSTM; Harini et al. (2025) fused TF-IDF with BERT embeddings; and T T et al. (2025) explored LLMs for detecting misogynistic slurs.

Our work adopts a different strategy: rather than modifying the architecture, we independently fine-tune an Indic-specialized and a general multilingual model, then combine their outputs through a calibrated probability-level ensemble.

3 Methodology

Given a Tamil social media comment x, we predict a label $y \in \{0, 1\}$, where $y = 1$ denotes `Abusive`

[1] https://github.com/meclin2345/AbuseDetect_Alchemists

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 144–147

and $y = 0$ denotes `Non-Abusive`.

3.1 Data Preparation

The organizers provided 3,652 labeled Tamil comments and 913 unlabeled test samples. We split the labeled data into 85% training (3,104 samples) and 15% validation (548 samples) using stratified sampling. Table 1 summarizes the splits.

Split	Non-Abusive	Abusive	Total
Train	1,600	1,504	3,104
Val	282	266	548
Test	–	–	913

Table 1: Dataset statistics across splits. The test set was provided unlabeled by the organizers; class-wise counts are therefore not available.

3.2 Text Preprocessing

We apply the following cleaning steps to all comments: (1) HTML entity decoding; (2) removal of user mentions (`@username`) and URLs; (3) whitespace normalization; and (4) emoji translation, where emoji characters are replaced with their CLDR short-name descriptors[2] (e.g., *face with tears of joy* → `:face_with_tears_of_joy:`). This last step is important because emojis frequently carry sarcastic or intensifying intent in abusive comments.

3.3 Model Selection and Fine-Tuning

We select two complementary transformers. **MuRIL** (Khanuja et al., 2021) is a BERT-based model pre-trained on 17 Indian languages including Tamil, making it effective at handling morphological complexity and code-mixed text. **XLM-RoBERTa** (Conneau et al., 2020) is pre-trained on 2.5 TB of CommonCrawl data covering 100 languages, offering broad cross-lingual coverage. Both models are loaded from the Hugging Face `transformers` library (Wolf et al., 2020) with a sequence classification head.

Each model is fine-tuned independently using cross-entropy loss with the hyperparameters in Table 2: AdamW optimizer (Loshchilov and Hutter, 2019), learning rate 2×10^{-5}, batch size 16, 5 epochs, max sequence length 128, weight decay 0.01, and linear warm-up scheduling.

[2]We use the Python `emoji` library (v2.12) for CLDR-based emoji-to-text conversion.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	16
Epochs	5
Max sequence length	128
Weight decay	0.01
LR scheduler	Linear with warm-up

Table 2: Fine-tuning hyperparameters.

3.4 Weighted Softmax Ensemble

For each test input x, let $\mathbf{z}^{(\mathrm{M})} \in \mathbb{R}^2$ and $\mathbf{z}^{(\mathrm{X})} \in \mathbb{R}^2$ denote the raw logit vectors from MuRIL and XLM-RoBERTa, respectively. The ensemble probability is computed as:

$$\mathbf{p}_{\text{ens}} = w_{\mathrm{M}} \cdot \text{softmax}(\mathbf{z}^{(\mathrm{M})}) + w_{\mathrm{X}} \cdot \text{softmax}(\mathbf{z}^{(\mathrm{X})}) \tag{1}$$

with $w_{\mathrm{M}} = 0.6$ and $w_{\mathrm{X}} = 0.4$. The higher weight for MuRIL reflects its stronger validation performance and closer alignment with Tamil. The final prediction is $\hat{y} = \arg\max_j \ \mathbf{p}_{\text{ens},j}$. This soft-voting strategy preserves confidence information, allowing a highly confident model to exert proportionally more influence.

4 Experiments

All experiments used the `transformers` (v4.44.2) (Wolf et al., 2020) and `datasets` (Lhoest et al., 2021) libraries on a single NVIDIA GPU. We used the model state after 5 epochs (970 steps per model), as both models showed stable convergence. The primary metric is **Macro-F1**, following the shared task guidelines.

5 Results and Analysis

5.1 Overall Performance

During fine-tuning, MuRIL achieved a validation Macro-F1 of 0.8011 by epoch 5, while XLM-RoBERTa peaked at 0.7973 (epoch 4) before a slight decline to 0.7954. Table 3 reports test set performance: both individual models achieve 0.8082 Macro-F1, while the weighted ensemble yields **0.8115**, a consistent improvement of 0.33 percentage points.

5.2 Class-wise Analysis

Table 4 shows per-class results on the test set. The ensemble achieves higher precision on Non-Abusive (0.8261) than Abusive (0.7969), while recall is marginally higher for Abusive (0.8186

Model	Val F1	Test F1
MuRIL	0.8011	0.8082
XLM-RoBERTa	0.7954	0.8082
Ensemble	**–**	**0.8115**

Table 3: Macro-F1 scores on validation and test sets.

Class	Precision	Recall	F1
Non-Abusive	0.8261	0.8051	0.8155
Abusive	0.7969	0.8186	0.8076
Macro Avg	0.8115	0.8118	**0.8115**

Table 4: Class-wise test set results for the weighted ensemble.

vs. 0.8051). The ensemble's predictions are near-balanced (460 Non-Abusive, 453 Abusive).

5.3 Error Analysis

Manual inspection of 50 misclassified samples reveals two recurring patterns. First, implicit and sarcastic abuse comments conveying hostility through sarcasm, rhetorical questions, or cultural innuendo without overt abusive vocabulary account for most false negatives. Second, heavy code-mixing between Tamil script, English, and romanized Tamil disrupts sub-word tokenization, causing fragmented token sequences and errors in both directions.

6 Conclusion

We presented a system for the DravidianLangTech@ACL 2026 shared task that combines noise-aware preprocessing with a weighted softmax ensemble of MuRIL (0.6) and XLM-RoBERTa (0.4), achieving a Macro-F1 of **0.8115**. The ensemble improves upon both individual models (0.8082) with balanced error reduction across classes. Future work will explore prompt-based LLMs for few-shot detection, learned ensemble weights via a meta-classifier, and data augmentation through back-translation to improve robustness to implicit abuse.

Limitations

Our work has several limitations. First, the ensemble weights ($w_{\mathrm{M}} = 0.6$, $w_{\mathrm{X}} = 0.4$) were set manually based on validation performance rather than learned through systematic optimisation (e.g., grid search or a meta-classifier), and may therefore be suboptimal for the test distribution. Second, our system treats the task as binary classification over individual comments and does not model conversational context; abusive intent that emerges across a thread of replies may be missed when each comment is classified in isolation. Third, the preprocessing pipeline translates emojis to English-language CLDR descriptors, which introduces a modality mismatch for models pre-trained primarily on Tamil text and may dilute the signal from Tamil-language tokens. Fourth, our error analysis is limited to a manual inspection of 50 misclassified samples, which is too small to draw statistically reliable conclusions about systematic failure modes.

References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John Philip McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

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.

Bachu Naga Sri Harini, Kankipati Venkata Meghana, Kondakindi Supriya, Tara Samiksha, and Premjith B. 2025. HTMS@DravidianLangTech 2025: Fusing TF-IDF and BERT with dimensionality reduction for abusive language detection in Tamil and Malayalam. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 152–156, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. 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, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual representations for Indian languages. *Preprint*, arXiv:2103.10730.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Maharajan Pannakkaran. 2025. byteSizedLLM@DravidianLangTech 2025: Abusive Tamil and Malayalam text targeting women on social media using XLM-RoBERTa and attention-BiLSTM. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 80–85, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

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*, pages 243–247, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan P, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 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, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Mirnalinee T T, J Bhuvana, Avaneesh Koushik, Diya Seshan, and Rohan R. 2025. SSNTrio@DravidianLangTech2025: LLM based techniques for detection of abusive text targeting women. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 415–419, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.