CHMOD_777@DravidianLangTech 2026: Context-Aware Fine-tuned MuRIL for Abusive Tamil Text Detection on Social Media

Arunaggiri Pandian Karunanidhi[1], **Prabalakshmi Arumugam**[2]
[1]Micron Technology, United States
[2]Boise State University, United States
arunaggiri.karunanidhi@gmail.com, prabalakshmiarumugam410@gmail.com

Abstract

This paper describes Team CHMOD_777's system for the DravidianLangTech@ACL 2026 shared task on detecting abusive Tamil text targeting women on social media. We fine-tune three transformer backbones (MuRIL (Multilingual Representations for Indian Languages), XLM-RoBERTa, IndicBERT-v3) with Focal Loss and weighted sampling, systematically evaluating the effects of context length, hyperparameter tuning, and language-specific pre-training. Our best system, MuRIL with 256-token context, achieves 82.76% Macro F1 on the development set and 80.61% on the official test set, ranking 6th out of 24 teams. We find that (1) extending context from 128 to 256 tokens improves F1 while converging 2.4× faster, (2) language-specific pre-training (MuRIL, 236M) outperforms larger models (IndicBERT, 270M), and (3) default hyperparameters are optimal, with every tuning attempt degrading performance.

1 Introduction

Online abuse targeting women on social media is a pervasive problem across languages and cultures. In low-resource languages such as Tamil, the challenge is amplified by code-mixing, transliteration, and culturally specific expressions of abuse. The DravidianLangTech@ACL 2026 shared task (Rajiakodi et al., 2026) addresses this by framing the detection of abusive Tamil text targeting women as a binary classification problem.

The task presents three challenges: (1) code-mixed Tamil-English text requiring multilingual tokenization; (2) transliteration where Tamil is written in Latin script on social media; and (3) culturally specific patterns of abuse with euphemisms that may not transfer from other languages.

Our approach is deliberately simple: fine-tuning pre-trained transformer models with Focal Loss (Lin et al., 2017) and weighted sampling, without custom architectural modifications. We focus instead on systematic evaluation of context length (128 vs. 256 tokens), model selection (language-specific vs. multilingual), and hyperparameter sensitivity. Our contributions are:

- MuRIL (Multilingual Representations for Indian Languages) with 256-token context achieves 82.76% Macro F1, ranking 6th out of 24 teams;
- Extending context from 128 to 256 tokens improves F1 while converging 2.4× faster;
- Language-specific pre-training outperforms larger general models, and default hyperparameters prove optimal.[1]

2 Related Work

Detecting abusive and hateful content on social media has received significant attention in NLP. Early approaches relied on handcrafted features and traditional machine learning classifiers (Waseem and Hovy, 2016), while recent work has shifted to transformer-based models that capture contextual nuances. For Indian languages specifically, shared tasks at DravidianLangTech workshops (Chakravarthi et al., 2022; Bharathi et al., 2022) have driven progress in code-mixed abuse detection and speech processing for Tamil, Telugu, and Malayalam, with transformer-based systems consistently outperforming feature-engineered baselines.

Tamil social media text poses unique challenges due to extensive code-mixing with English and frequent transliteration from Tamil script to Latin characters. Jada et al. (2021) demonstrated that transformer architectures significantly outperform traditional methods for Dravidian language sentiment analysis. Sai Kumar et al. (2021) showed that

[1]The code for this work is available at: https://github.com/Arunaggiri-Pandian/Abusive_Tamil_Text_Classification

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 176–180

Split	Non-Abusive	Abusive	Total	Abusive %
Train	1,694	1,592	3,286	48.4
Dev	189	177	366	48.4
Test	472	441	913	48.3

Table 1: Dataset statistics. Nearly balanced (1:1 ratio).

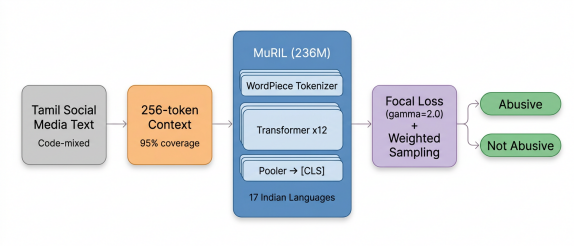

Figure 1: System pipeline: input text tokenized to 256 tokens, encoded by MuRIL, classified with Focal Loss.

BERT-based models achieve competitive performance for tweet-level classification tasks involving code-mixed text. Prior work on multimodal Dravidian content analysis (Premjith et al., 2022) has further established that pre-trained language models can capture cross-lingual patterns in mixed-script settings.

Among pre-trained models, MuRIL (Khanuja et al., 2021), pre-trained on 17 Indian languages with transliteration augmentation, has shown strong results for Tamil. XLM-RoBERTa (Conneau et al., 2020) provides broad multilingual coverage but dilutes language-specific representations across 100+ languages. IndicBERT-v3 (Doddapaneni et al., 2023) targets Indian languages specifically but with a different pre-training strategy. Our work provides a direct comparison of these three paradigms on the same task and dataset.

Focal Loss (Lin et al., 2017), originally proposed for class-imbalanced object detection, down-weights well-classified examples through a focusing parameter γ and concentrates learning on hard examples near the decision boundary. While primarily used for imbalanced datasets, we demonstrate its utility even for nearly balanced binary classification where ambiguous cases benefit from focused training.

3 Methodology

3.1 Dataset

The dataset comprises 3,286 training, 366 development, and 913 test samples. Table 1 shows the class distribution. The dataset is nearly balanced (1:1 ratio) with identical proportions across train and dev splits, indicating careful stratified splitting.

Tamil social media text presents specific challenges. Code-mixing is highly prevalent, with tweets frequently alternating between Tamil and English within a single utterance. Transliteration adds further complexity as Tamil is often written using Latin characters (e.g., "romba bayama iruku" instead of native Tamil script). The average text length is approximately 111 characters (14 words), though abusive content tends to use longer expressions with multiple clauses.

3.2 Model Architecture

We adopt a straightforward classification approach: each pre-trained encoder maps an input text to a pooled `[CLS]` representation $\mathbf{h} \in R^d$, which is classified via a dropout layer (p=0.1) and a linear head $W \in R^{d \times 2}$ (Devlin et al., 2019). We compare three backbones that represent different pre-training strategies, all sharing the same hidden dimension (d=768) and 12-layer architecture:

- **MuRIL** (Khanuja et al., 2021): 236M params, pre-trained on 17 Indian languages with transliteration augmentation, providing native Tamil coverage including romanized text.
- **XLM-RoBERTa** (Conneau et al., 2020): 278M params, pre-trained on 100+ languages via CommonCrawl, offering broad but diluted per-language coverage.
- **IndicBERT-v3** (Doddapaneni et al., 2023): 270M params, pre-trained on 11 Indic languages with IndicCorp data, using a different pre-training strategy than MuRIL.

All models are fine-tuned end-to-end using HuggingFace Transformers. A key design decision is the input context length: we evaluate 128 and 256 tokens to study how context affects Tamil abuse detection. We chose 128 as the standard BERT default and 256 as a practical upper bound that covers 95% of samples in the dataset without excessive padding. Longer contexts (384, 512) would introduce majority-padding tokens that dilute attention, particularly problematic for the shorter social media texts in this corpus. Figure 1 shows the overall pipeline.

3.3 Training Configuration

We use Focal Loss (Lin et al., 2017) with γ=2.0 and inverse-frequency class weights α_t:

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (1)$$

Parameter	Value
Max sequence length	256 (or 128 for ablation)
Batch size	16
Learning rate	1e-5
Optimizer	AdamW
Weight decay	0.01
Loss function	Focal (γ=2.0)
Gradient clipping	max_norm = 1.0
Early stopping patience	10 epochs
Max epochs	50
Seed	42

Table 2: Training hyperparameters (MuRIL, 256 tokens).

Model	Tokens	MF1	Epoch	Δ
MuRIL v2	**256**	**82.76**	**9**	
MuRIL v1	128	82.50	22	−0.26
MuRIL Tuned	128	82.45	22	−0.31
MuRIL Tuned+GA	128	82.23	17	−0.53
XLM-RoBERTa	256	81.95	10	−0.81
IndicBERT-v3	256	74.02	39	−8.74

Table 3: Dev set results. Tuned = lr 2e-5, warmup, γ=1.0. GA = gradient accumulation.

where p_t is the predicted probability for the true class. The focusing term $(1 - p_t)^\gamma$ down-weights well-classified examples, concentrating learning on hard examples near the decision boundary. Even though the dataset is nearly balanced, this is valuable for abuse detection where the distinction between abusive and non-abusive content is often subjective. WeightedRandomSampler ensures balanced mini-batches, and early stopping monitors dev Macro F1 with patience of 10 epochs. Table 2 lists the full configuration.

To assess hyperparameter sensitivity, we train additional MuRIL variants with increased learning rate (2e-5), linear warmup (10%), higher weight decay (0.02), reduced focal gamma (1.0), and gradient accumulation (effective batch size 32). These experiments allow us to isolate the impact of each training decision.

4 Results

4.1 Development Set Performance

Table 3 presents all six experiments. MuRIL with 256-token context achieves the best Macro F1 of 82.76%, converging at epoch 9.

4.2 Per-Class Analysis

Table 4 shows per-class metrics for the best model. The F1 gap between classes is only 1.41 percentage points, demonstrating symmetric performance.

Class	Precision	Recall	F1
Non-Abusive	82.81	84.13	83.46
Abusive	82.76	81.36	82.05
Macro Avg	**82.79**	**82.74**	**82.76**

Table 4: Per-class metrics (%) for MuRIL v2 (256 tokens).

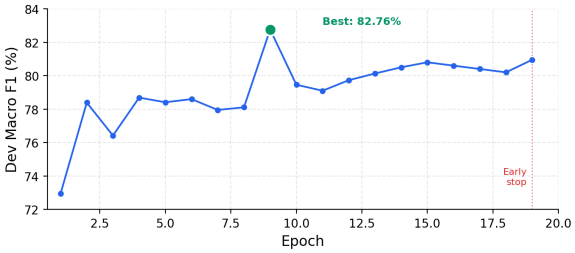

Figure 2: Training progression. F1 peaks at epoch 9.

The largest asymmetry is in recall (2.77 point gap), where the model is slightly better at identifying Non-Abusive text, likely reflecting the inherent difficulty of detecting subtler forms of abuse.

4.3 Training Dynamics

Figure 2 shows the training progression for MuRIL v2. The model exhibits rapid initial learning (F1 jumps from 72.95% to 78.39% in one epoch), peaks at epoch 9 (82.76%), and remains stable at 80.95% by epoch 19 when early stopping triggers. MuRIL v2 retains within 1.81 points of its peak over 10 additional epochs, indicating effective regularization from Focal Loss and the balanced sampler.

4.4 Official Test Set Results

On the official test set, our primary submission achieves 80.61% Macro F1, ranking 6th out of 24 teams. Table 5 compares dev and test per-class metrics. The symmetric performance holds on the test set, with only 0.68% F1 gap between classes (Non-Abusive 80.95%, Abusive 80.27%). The competition was highly competitive, with the top 10 spanning only 2.96 points. Full leaderboard results are available in the task overview paper (Rajiakodi et al., 2026).

5 Analysis

Extending context from 128 to 256 tokens yields the most impactful improvement: MuRIL (256 tokens, F1=82.76%) outperforms its 128-token variant (F1=82.50%) while converging 2.4× faster (9 vs 22 epochs), as shown in Figure 3. The extended

Class	Dev F1	Test P	Test R	Test F1
Non-Abusive	83.46	82.28	79.66	80.95
Abusive	82.05	78.95	81.63	80.27
Macro	**82.76**	**80.61**	**80.65**	**80.61**

Table 5: Per-class dev vs test (%). P = Precision, R = Recall.

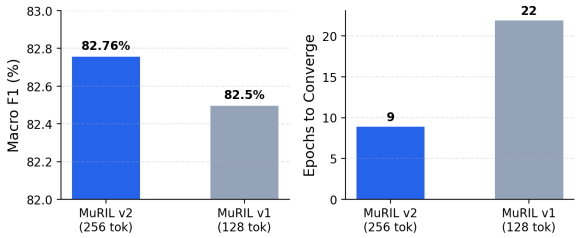

Figure 3: Context length ablation: 256 tokens achieve higher F1 with 2.4× faster convergence.

context captures complete code-mixed expressions that frequently exceed 128 tokens.

MuRIL (236M, F1=82.76%) outperforms XLM-RoBERTa (278M, 81.95%) and IndicBERT-v3 (270M, 74.02%) despite being smallest, confirming that pre-training strategy matters more than scale. Every hyperparameter tuning attempt degraded performance (largest: gradient accumulation −0.53%), as the small dataset (3,286 samples) means pre-trained weights already provide a well-conditioned loss landscape where conservative defaults are optimal. Figure 4 compares all backbones.

Error analysis reveals two failure modes: (1) subtle abuse conveyed through sarcasm or cultural euphemisms without explicit keywords, and (2) code-mixed texts where the abusive signal spans Tamil and English fragments. Focal Loss produces symmetric per-class performance (1.41% F1 gap) despite near-balanced data, confirming its value for focusing on ambiguous boundary cases.

6 Conclusion

We presented a system for detecting abusive Tamil text achieving 80.61% Macro F1 (Rank 6/24). Key findings: (1) 256-token context improves F1 and converges 2.4× faster; (2) MuRIL outperforms larger models by up to 8.74 points; (3) default hyperparameters are optimal on small datasets; (4) Focal Loss benefits even balanced classification.

Limitations

Data augmentation techniques tailored to Tamil text could yield further improvements. Ensemble

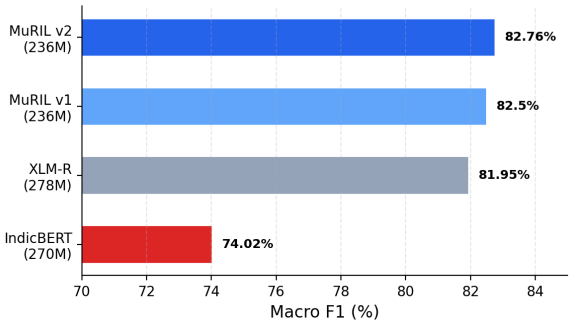

Figure 4: Model comparison. MuRIL achieves the best F1 despite fewest parameters.

methods combining MuRIL with XLM-RoBERTa may leverage their complementary strengths across different text types. Incorporating external lexical resources for Tamil abusive language is another avenue that could improve recall on subtle forms of abuse.

References

B. Bharathi, Bharathi Raja Chakravarthi, Subalalitha Cn, N. Sripriya, Arunaggiri Pandian Karunanidhi, and Swetha Valli. 2022. Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Rahul Ponnusamy, John P. McCrae, Elizabeth Sherly, and Saranya Rajiakodi. 2022. Findings of the shared task on sentiment analysis in Tamil and Telugu code-mixed text. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. 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*. 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 (NAACL-HLT)*. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Divyanshu Kakwani, Anoop Kumar, Jessica Murber, Preethi

Nemani, and Ravi Shankar Lakumarapu. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. *arXiv preprint arXiv:2212.05409*.

Pawan Kalyan Jada, D. Sashidhar Reddy, Konthala Yasaswini, Arunaggiri Pandian Karunanidhi, Prabakaran Chandran, Anbukkarasi Sampath, and Sathiyaraj Thangasamy. 2021. Transformer based sentiment analysis in Dravidian languages. In *Proceedings of FIRE 2021 (Working Notes)*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagisetty, Shachi Gupta, Ankur Parikh, and Partha Talukdar. 2021. MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

B. Premjith, Bharathi Raja Chakravarthi, Malliga Subramanian, B. Bharathi, K.P. Soman, V. Dhanalakshmi, K. Sreelakshmi, Arunaggiri Pandian Karunanidhi, and Prasanna Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. 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.

T.S. Sai Kumar, K. Arunaggiri Pandian Karunanidhi, S. Thabasum Aara, and K. Nagendra Pandian. 2021. A reliable technique for sentiment analysis on tweets via machine learning and BERT. In *2021 Asian Conference on Innovation in Technology (ASIANCON)*. IEEE.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics.