PrimeLine@DravidianLangTech 2026: Abusive Tamil Comment Detection Using MuRIL

Rithikaa V **Sanjay Krishnan K** **Nithya Varshini C N R** **S. Sumathi**
Department of Information Technology, St. Joseph's College of Engineering
https://github.com/Nithya-svg/primeline_tamil_abusive

Abstract

Detecting abusive language in Tamil social media is a genuinely difficult problem. The language is morphologically rich, speakers routinely mix Tamil with English, and informal romanised Tamil is common enough to confuse models trained primarily on formal text. This work presents a system for binary classification of Tamil comments into Abusive and Non-Abusive categories, submitted to the DravidianLangTech@ACL 2026 shared task [12]. MuRIL [5], a BERT-based encoder pre-trained on 17 Indian languages and their transliterated equivalents, is fine-tuned, and it is shown that this Indian-language-specific pre-training provides a meaningful advantage over generic multilingual baselines. The system achieves a macro-averaged F1 of 0.83 on the validation set, compared to 0.79 for XLM-RoBERTa and 0.77 for mBERT under identical training conditions, establishing a strong transformer-based baseline for abusive language detection in code-mixed Tamil.

1 Introduction

Tamil is one of the oldest classical languages still in everyday use, with more than 80 million speakers worldwide. Social media has given Tamil speakers an enormous platform for public expression, but it has also made it easier for abusive and harmful content to spread. Detecting such content automatically is a growing research priority, yet most existing systems are designed for high-resource languages like English and do not transfer well to Tamil.

The challenge goes beyond data scarcity. Tamil is an agglutinative language, where a single word can encode morphological information that would require several words in English. On social media, this complexity is compounded by heavy code-mixing: users blend native Tamil script, romanised Tamil, and English freely within a single comment, with no consistent orthographic convention. Lexicon-based systems and classical machine learning approaches struggle in this setting because they rely on vocabulary stability that code-mixed text simply does not have.

The DravidianLangTech@ACL 2026 shared task on abusive Tamil comment detection [12] provides a structured benchmark for this problem, framing it as binary classification: given a Tamil social media comment, decide whether it is Abusive or Non-Abusive. The approach centers on fine-tuning MuRIL [5], a transformer encoder pre-trained specifically on 17 Indian languages including Tamil, using both native script and transliterated corpora. MuRIL's dual-script coverage makes it uniquely well-suited to the romanised Tamil that is pervasive in online communication.

2 Related Work

Early systems for abusive language detection relied on keyword lists and hand-crafted features combined with traditional classifiers such as Support Vector Machines and Naive Bayes [8]. While effective for formal single-language text, these methods break down when vocabulary is informal or code-mixed.

The introduction of BERT [4] changed the landscape considerably. Pre-trained transformer models produced contextual representations that generalised far better across domains, and their multilingual variants extended these benefits to non-English languages. Ranasinghe and Zampieri [9] demonstrated that fine-tuned multilingual transformers consistently outperform traditional methods on offensive language tasks across multiple languages. Vaswani et al. [14] introduced the Transformer architecture that underpins all of these models, establishing multi-head self-attention as the mechanism for capturing long-range token dependencies. Liu et al. [7] later showed that more robust pre-training strategies yield better downstream

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 331–335

performance, a finding extended to the multilingual setting by Conneau et al. [3] through XLM-RoBERTa.

For Dravidian languages, the DravidianLangTech and LT-EDI shared tasks have been the primary source of benchmarks. Chakravarthi et al. [2] reported results on offensive language identification in Tamil, Malayalam, and Kannada, where fine-tuned mBERT and XLM-RoBERTa were the dominant approaches. Sai and Sharma [11] explored offensive language identification across Dravidian languages. Tamil-specific work has explored ensemble strategies and model combinations: Anbukkarasi and Varadhaganapathy [1] applied deep learning for hate speech detection in code-mixed Tamil, Subramanian et al. [13] used adapters and cross-domain transfer for offensive language in Tamil YouTube comments, and Krishna et al. [6] addressed class imbalance in Tamil code-mixed abusive comment detection. S et al. [10] applied RoBERTa and XGBoost for abusive Tamil and Malayalam text detection at DravidianLangTech 2025. However, MuRIL has not been systematically evaluated on abusive language detection in Tamil — which is the gap this work addresses.

MuRIL [5] was developed by Google Research and pre-trained on 17 Indian languages using Wikipedia and Common Crawl data in both native and transliterated script. This dual-script design is a key advantage over XLM-RoBERTa and mBERT, neither of which includes transliterated training data. For Tamil social media, where romanised text is widespread, MuRIL's pre-training distribution is a much closer match to the target domain.

3 Dataset and Preprocessing

3.1 Dataset

The dataset was provided by the shared task organisers and consists of Tamil social media comments labelled as Abusive or Non-Abusive. The text reflects real-world Tamil online writing, freely mixing native Tamil script, romanised Tamil, and English within individual comments. This code-mixed, multi-script nature makes automated classification considerably harder than a standard monolingual setting.

The full annotated training set contains 3,652 examples. A stratified 90/10 split is applied to preserve label distribution, yielding 3,286 training examples and 366 validation examples. The test set comprises 912 unlabelled comments for blind evaluation, which is standard practice in shared task settings. The training data exhibits moderate class imbalance: approximately 64% Non-Abusive and 36% Abusive. Table 1 shows the full class distribution. Comment lengths vary from brief single-word utterances to multi-sentence posts, with a mean token length of approximately 18 tokens after Unicode normalisation. The minority Abusive class is the harder of the two to detect reliably, especially when abuse is expressed through sarcasm or indirect phrasing rather than explicit offensive vocabulary.

Label	Count	%
Non-Abusive	2338	64%
Abusive	1314	36%
Total	3652	100%

Table 1: Class distribution for the Tamil abusive comment detection training data.

3.2 Preprocessing

Text samples are tokenised using the MuRIL WordPiece tokeniser, which handles both native Tamil script and romanised equivalents natively. A maximum sequence length of 128 tokens is used, sufficient for the short social media comments in this dataset. Sequences exceeding this limit are truncated, and shorter sequences are padded within each batch. Unicode normalisation is applied to handle encoding inconsistencies common in crawled social media text.

Transliteration, language identification, and stopword removal are deliberately skipped. MuRIL's dual-script pre-training already accounts for the mix of native and romanised Tamil, and additional normalisation risks erasing code-switching cues that carry genuine semantic meaning.

4 Methodology

4.1 System Pipeline

The system follows a straightforward fine-tuning pipeline, illustrated in Figure 1. Each comment is passed to the MuRIL tokeniser, encoded as token IDs with a prepended `[CLS]` token, and fed into the MuRIL transformer encoder. The hidden state at the `[CLS]` position is extracted and passed through a linear classification head that maps it to two logits. These are normalised via softmax, and the class with the highest probability is taken as the prediction.

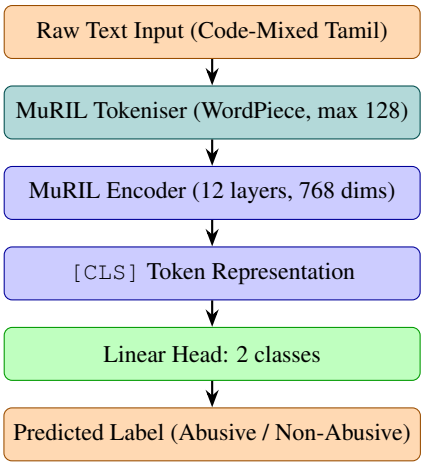

Figure 1: End-to-end system pipeline for abusive comment classification in Tamil.

4.2 Model: MuRIL

The core model is MuRIL (Multilingual Representations for Indian Languages; Khanuja et al. 5), a BERT-based encoder released by Google Research, pre-trained on 17 Indian languages in both native and transliterated script. Its 12-layer transformer architecture produces 768-dimensional contextualised token representations. A single linear classification head is placed over the `[CLS]` token representation, mapping the 768-dimensional vector to two output logits. The entire model is fine-tuned end-to-end.

4.3 Model Formulation

At each transformer layer, MuRIL computes scaled dot-product attention over queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \quad (1)$$

where d_k is the key vector dimension. This allows the model to capture dependencies between tokens across Tamil script, romanised Tamil, and English simultaneously within a single comment.

The `[CLS]` representation from the final encoder layer is passed through the linear head, and the resulting logits are normalised via softmax to produce class probabilities for $c \in \{\text{Abusive, Non-Abusive}\}$:

$$P(y=c \mid x) = \frac{\exp(z_c)}{\sum_j \exp(z_j)} \quad (2)$$

Training minimises binary cross-entropy loss over all training samples:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log P(y{=}1 \mid x_i) + (1-y_i)\log P \quad (3$$

where N is the number of training examples and $y_i \in \{0,1\}$ is the binary ground truth label.

4.4 Training Setup

All models are fine-tuned using the Hugging Face Transformers Trainer API [15] with a learning rate of 2×10^{-5}, batch size of 16, weight decay of 0.01, and 4 training epochs. The AdamW optimiser is used throughout. Macro-averaged F1 on the validation set serves as the primary evaluation metric, since it weights all classes equally and is sensitive to performance on the minority Abusive class. All experiments run on a CUDA-enabled GPU with total training time of approximately 25 minutes.

MuRIL was evaluated at 2, 3, 4, and 5 training epochs on the validation set. Performance improved from 0.79 at 2 epochs to 0.83 at 4 epochs, and stabilised thereafter with no further gain at 5 epochs. The 4-epoch configuration was therefore used for all reported experiments.

5 Results and Discussion

5.1 Official Results

Table 2 reports macro-averaged precision, recall, and F1 on the held-out validation set for MuRIL and two multilingual baselines — mBERT [4] and XLM-R [3] — both fine-tuned under identical settings.

Model	P	R	F1
mBERT	0.78	0.76	0.77
XLM-R	0.80	0.79	0.79
MuRIL (proposed)	**0.84**	**0.83**	**0.83**

Table 2: Macro-averaged results on the validation set (P = Precision, R = Recall, F1 = Macro F1).

MuRIL achieves the best performance across all metrics, with a macro-F1 of 0.83 compared to 0.79 for XLM-R and 0.77 for mBERT. The improvement is most pronounced for the Abusive class, where MuRIL's Indian-language-specific pre-training enables better representation of Tamil morphology and colloquial script variants. The gap between MuRIL and XLM-R is especially telling: XLM-R is a strong multilingual model, but it does not include transliterated training data. For Tamil social

media, that gap in pre-training coverage translates directly into lower recall on abusive content.

Table 3 shows the per-class breakdown for MuRIL, giving a clearer view of where the model performs well and where it still falls short.

Class	Precision	Recall
Non-Abusive	0.87	0.91
Abusive	0.81	0.75

Table 3: Per-class precision and recall for the proposed MuRIL system.

The Non-Abusive class is handled with higher precision and recall, as expected given the class imbalance. A recall of 0.75 on Abusive comments means roughly one in four abusive examples is missed — improving this is the clearest direction for future work.

5.2 Error Analysis

Qualitative examination of the misclassified examples reveals two dominant patterns. The first is implicit abuse: comments that convey abusive intent through sarcasm, euphemism, or cultural reference without any overtly offensive vocabulary. These are consistently misclassified as Non-Abusive by all three models, since the meaning only becomes clear with contextual knowledge the models do not possess.

The second pattern involves dense code-mixing: utterances that switch between Tamil script, romanised Tamil, and English multiple times within a single short comment. When all three layers are simultaneously active, even MuRIL occasionally fails to build a coherent contextual representation.

False negatives (Abusive predicted as Non-Abusive) outnumber false positives across all models, reflecting the majority-class bias introduced by the training imbalance. Future work should explore loss reweighting or focal loss to counteract this tendency, as well as back-translation from related Dravidian languages to augment the Abusive training set.

6 Conclusion

This work presents a system for Tamil abusive comment detection based on fine-tuning MuRIL for binary sequence classification. By leveraging Indian-language-specific pre-training on both native and transliterated scripts, the model achieves a macro-F1 of 0.83 on the validation set, outperforming XLM-R and mBERT by clear margins under identical conditions. The results confirm that pre-training data composition matters: a model exposed to transliterated Tamil is better equipped to handle the romanised code-mixing that defines Tamil social media. Future directions include adversarial training for implicit abuse, ensemble methods combining transformer representations with lexicon-based features, and extension to multi-class offensive language categories.

Limitations

The system is trained and evaluated on the provided shared task dataset alone, and generalisation to other Tamil platforms or time periods cannot be guaranteed. The maximum sequence length of 128 tokens may truncate longer comments, losing relevant context. Computational constraints limited experiments to MuRIL-base. Class imbalance was not addressed through resampling or weighted loss, which likely suppressed recall on the Abusive class.

Acknowledgments

The organizers of the DravidianLangTech@ACL 2026 shared task [12] are thanked for providing the annotated dataset and evaluation infrastructure. Sincere gratitude is expressed to the project guide, S. Sumathi, Department of Information Technology, St. Joseph's College of Engineering, for constant guidance and encouragement throughout this work. The developers of MuRIL and the Hugging Face Transformers library are also thanked for making their tools openly available.

References

[1] S. Anbukkarasi and S. Varadhaganapathy. Deep learning-based hate speech detection in code-mixed Tamil text. *IETE Journal of Research*, 69(11):7893–7898, 2023. URL https://doi.org/10.1080/03772063.2021.1992853.

[2] Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Navaneethan Rajasekaran, Sajeetha Thavareesan, and Dhivya Chinnappa. 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, 2021. URL https://aclanthology.org/2021.dravidianlangtech-1.17.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer,

and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020. URL https://aclanthology.org/2020.acl-main.747.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, 2019. URL https://aclanthology.org/N19-1423.

[5] 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. MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*, 2021. URL https://arxiv.org/abs/2103.10730.

[6] R. Krishna et al. Abusive comment detection in Tamil code-mixed data by adjusting data distribution. In *Proceedings of the International Conference on Computational Intelligence and Data Engineering (ICCIDE 2024)*, 2024. URL https://doi.org/10.1007/978-981-97-6714-4.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

[8] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153, 2016. URL https://doi.org/10.1145/2872427.2883062.

[9] Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5838–5847, 2021. URL https://aclanthology.org/2021.emnlp-main.470.

[10] Nishanth S, Shruthi Rengarajan, S Anantha-sivan, Burugu Rahul, and Sachin Kumar S. ANSR@DravidianLangTech 2025: Detection of abusive Tamil and Malayalam text targeting women on social media using RoBERTa and XGBoost. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 711–715, 2025. URL https://aclanthology.org/2025.dravidianlangtech-1.98.

[11] Siva Sai and Yashvardhan Sharma. Towards offensive language identification for Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 18–27, 2021. URL https://aclanthology.org/2021.dravidianlangtech-1.3.

[12] Bhuvaneswari Sivagnanam, Kathiravan Pannerselvam, Jananayagan V, Charmathi Rajkumar, Ramesh Kannan R, Ratnavel Rajalakshmi, Shunmuga Priya Muthusamy Chinnan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 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, 2026. URL https://aclanthology.org/2026.dravidianlangtech.

[13] S. Subramanian et al. Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 74:101386, 2022. URL https://doi.org/10.1016/j.csl.2022.101386.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-html.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and Jamie Brew. 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, 2020. URL https://aclanthology.org/2020.emnlp-demos.6.