PhucNguyen@DravidianLangTech 2026: Political Multiclass Sentiment Analysis with XLM-RoBERTa and Low-Rank Adaptation

Dinh Khac Phuc Nguyen[1,2] **and Dang Van Thin**[1,2]
[1]University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
24521172@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

Analyzing political sentiment in code-mixed Tamil-English presents significant challenges due to informal jargon, severe class imbalance, and distribution shifts. This paper describes our system for the Political Multiclass Sentiment Analysis shared task at DravidianLangTech@ACL 2026, which categorizes tweets into seven sentiment classes. Our approach leverages XLM-RoBERTa integrated with Low-Rank Adaptation (LoRA). To mitigate majority-class dominance, we combine random oversampling with automated hyperparameter optimization to improve macro-level balance within this Parameter-Efficient Fine-Tuning (PEFT) framework. Enhanced by targeted preprocessing—specifically emoji demojization and noise removal—our system helps preserve nuanced symbolic cues, achieving a macro-average F1-score of 0.3763 and securing Rank 2 on the shared task leaderboard.

Keywords: Sentiment Analysis, Code-Mixed Data, Tamil tweets, XLM-RoBERTa, PEFT Tuning.

1 Introduction

Analyzing political sentiment in low-resource, code-mixed languages like Tamil-English is highly challenging due to noisy user-generated content and severe class imbalance, which frequently causes neural models to exhibit systematic majority-class bias (Chakravarthi et al., 2020). While multilingual models such as XLM-R (Conneau et al., 2019) and Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA (Hu et al., 2022) offer robust and efficient solutions, their behavior under extreme class skew remains empirically under-characterized in shared-task settings. Specifically, we observe that default learning rate configurations in LoRA can rapidly exacerbate majority-class dominance, resulting in severe prediction bias toward dominant categories.

In this work, we investigate LoRA-based fine-tuning for Political Multiclass Sentiment Analysis of Tamil-English tweets (Vegupatti et al., 2026). By coupling random oversampling of the training data with a systematic hyperparameter search via Optuna, we demonstrate that exploring learning rate and weight decay sensitivity effectively mitigates majority-class over-indexing and improves macro-level balance. Furthermore, we show that domain-specific preprocessing—namely emoji demojization and hashtag normalization—helps preserve essential symbolic and lexical cues in informal text.

Our contributions are threefold:

- We detail an optimized LoRA adaptation pipeline under severe class imbalance, highlighting the model's sensitivity to hyperparameter configurations and the practical necessity of automated tuning.
- We demonstrate that targeted preprocessing (demojization and hashtag segmentation) effectively retains semantic signals in code-mixed political discourse.
- Our system achieves a macro-F1 score of 0.3763—an absolute improvement of +0.0331 over a standard full fine-tuning baseline—securing Rank 2 in the DravidianLangTech 2026 shared task (Vegupatti et al., 2026).

The remainder of this paper is organized as follows. Section 2 reviews related work relevant to our study. Section 3 presents the task and dataset description, including the problem formulation and data characteristics. Section 4 details the proposed methodology. Section 5 describes the experimental setup and reports the results. Section 6 provides discussion and in-depth analysis of the findings. Finally, Section 7 concludes the paper and outlines directions for future work.

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 321–325

2 Related Work

Code-Mixed Sentiment Analysis. Sentiment analysis in Tamil-English (Tanglish) is hindered by informal grammar, phonetic typing (Chakravarthi et al., 2020; Hande et al., 2021), and symbolic noise (e.g., emojis), which standard subword tokenizers inconsistently segment (Priyadharshini et al., 2021). Furthermore, Dravidian political sentiment datasets exhibit severe skew toward opinionated content, exacerbating minority class sparsity. This necessitates domain-specific text normalization to preserve contextual cues prior to encoding.

Multilingual Models and PEFT. While multilingual models like XLM-RoBERTa (Conneau et al., 2019) excel in cross-lingual transfer, full fine-tuning on small, skewed datasets frequently causes unstable adaptation and overfitting. Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA (Hu et al., 2022) mitigate this by restricting the trainable parameter space. However, under extreme class imbalance, this reduced capacity may amplify sensitivity to hyperparameter configurations, leaving PEFT optimization behavior less explored in shared-task contexts.

Optimization under Class Imbalance. Traditional imbalance mitigations like Focal Loss (Lin et al., 2017) or data resampling can introduce instability or increase overfitting risks under noisy supervision. To manage these trade-offs, automated hyperparameter optimization (Akiba et al., 2019) is utilized. Unlike prior work focusing solely on objective re-weighting, we emphasize the interaction between data-level and optimization-level interventions. We empirically explore how coupling data resampling with automated tuning within the constrained LoRA space improves training stability and mitigates majority-class dominance, offering a practical alternative to bespoke loss engineering.

3 Task and Dataset Description

3.1 Task Definition and Evaluation

The DravidianLangTech@ACL 2026 shared task (Vegupatti et al., 2026) requires classifying code-mixed Tamil tweets into seven political sentiment classes. Systems are evaluated using the macro-averaged F1-score. By assigning equal weight to each class, this metric strictly emphasizes minority-class performance, making robustness to imbalance crucial.

Table 1: Class distribution of the official datasets. The severe skew toward *Opinionated* presents a significant optimization challenge.

Sentiment Class	Train (%)	Dev (%)
Opinionated	1,361 (31.3)	153 (28.1)
Sarcastic	790 (18.2)	115 (21.1)
Neutral	637 (14.6)	84 (15.4)
Positive	575 (13.2)	69 (12.7)
Substantiated	412 (9.5)	52 (9.6)
Negative	406 (9.3)	51 (9.4)
None of the above	171 (3.9)	20 (3.7)
Total	**4,352 (100)**	**544 (100)**

3.2 Dataset Statistics and Optimization Challenges

The Tanglish dataset includes 4,352 training and 544 development instances, featuring informal text rich in political hashtags and emojis. The class distribution remains consistent across both splits.

As shown in Table 1, the data is severely skewed: *Opinionated* dominates (31.3%), while *Substantiated* and *Negative* fall below 10%. In practice, full fine-tuning under such imbalance often yields majority-class-biased predictions, as frequent labels dominate the optimization objective. We hypothesize that LoRA's restricted trainable parameter space influences adaptation dynamics under skewed supervision, potentially moderating abrupt majority-class bias compared to full-parameter updates.

4 Methodology

Our system integrates contextual text normalization, cross-lingual encoding, and parameter-efficient optimization to address the vulnerabilities of pretrained models under extreme class skew and phonetic noise.

4.1 Contextual Preprocessing

Political discourse relies heavily on symbolic cues frequently inconsistently segmented by subword tokenizers. We implemented targeted preprocessing to anchor these within the embedding space: (1) **Emoji Demojization** via the `emoji` library to translate symbols into descriptive text, and (2) **Structural Noise Removal** filtering URLs and user mentions. Additionally, sequences were padded and truncated to a maximum length of 256 tokens to balance contextual coverage and computational efficiency.

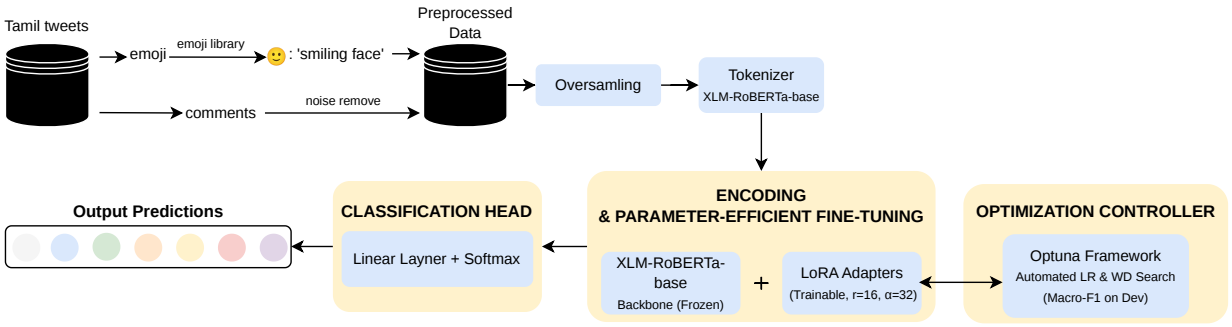

Figure 1: Overview of the XLM-R + LoRA pipeline with Optuna-based hyperparameter optimization for robust training under class imbalance and noisy social media text.

4.2 Multilingual Foundation Model

We adopted **XLM-RoBERTa-base** (Conneau et al., 2019), leveraging its robust cross-lingual representations for Tanglish text without requiring external translation pipelines.

4.3 Parameter-Efficient Adaptation via LoRA

To prevent severe overfitting on small, skewed datasets, we applied **Low-Rank Adaptation (LoRA)** (Hu et al., 2022). LoRA freezes pre-trained weights and introduces trainable low-rank matrices into the self-attention layers. We targeted the *query* and *value* projections ($r = 16$, $\alpha = 32$, dropout=0.05), reducing trainable parameters to under 1%, thereby limiting adaptation capacity and improving parameter efficiency. All configurations utilized mixed-precision (FP16) training to optimize memory utilization.

4.4 Imbalance Mitigation via Optimization

To alleviate initial data sparsity, we applied random oversampling (Lemaître et al., 2017) exclusively to the training set. However, naive oversampling risks overfitting to duplicated minority-class noise if the learning process is not carefully controlled.

To systematically navigate the constrained LoRA parameter space, we utilized **Optuna** (Akiba et al., 2019) for hyperparameter search. Empirical tuning identified an optimal hyperparameter configuration—specifically a learning rate of 1.31×10^{-4} (AdamW) and a weight decay of 0.098—providing stable adaptation. This hybrid strategy effectively improves minority-class performance without the need for bespoke loss engineering.

Table 2: Hyperparameter optimization details explored via Optuna.

Hyperparameter	**Search Space**	**Best Value**
Learning Rate	$[1 \times 10^{-4}, 5 \times 10^{-4}]$	1.31×10^{-4}
Weight Decay	$[0.01, 0.1]$	0.098
Num. Train Epochs	$[6, 15]$	10
AdamW β_1	$[0.8, 0.99]$	0.8
AdamW ϵ	$[10^{-8}, 5 \times 10^{-6}]$	6×10^{-8}
Fixed Parameters		
LoRA Rank (r)	16	16
LoRA Alpha (α)	32	32
Batch Size	32	32

Table 3: System performance on the test set.

Model Configuration	**Accuracy**	**Macro F1**
XLM-R (Full Fine-Tuning)	0.3033	0.3432
XLM-R + LoRA (Standard)	0.3033	0.3193
XLM-R + LoRA + Focal Loss	0.3511	0.3286
XLM-R + LoRA + Optuna (Ours)	**0.3364**	**0.3763**

5 Experimental Setup and Results

5.1 Evaluation Metrics

Following shared task guidelines, we evaluate system performance using the macro-averaged F1-score to ensure smaller sentiment categories are appropriately weighted. This prevents the predominant *Opinionated* class from disproportionately skewing the perceived efficacy. Overall accuracy is also reported for completeness.

5.2 Comparative Results

Table 3 details the performance across configurations. Our final pipeline—integrating XLM-RoBERTa, LoRA, and Optuna—achieved a macro-F1 of 0.3763 and an accuracy of 0.3364 on the official blind test set, securing 2nd place in the shared task.

5.3 Analysis

The results highlight the complexities of adaptation under extreme class skew. Standard LoRA underperforms the full fine-tuning baseline in macro-F1 (0.3193 vs. 0.3432). Notably, while Focal Loss achieves the highest overall accuracy (0.3511), its lower macro-F1 (0.3286) suggests less balanced performance across sentiment categories, as evidenced by the discrepancy between overall accuracy and macro-averaged metrics.

Conversely, systematic hyperparameter search via Optuna yielded a notable improvement of **0.0331 macro-F1 points** over the baseline. The contrast between Focal Loss and Optuna-based tuning highlights an important trade-off between objective-level re-weighting and optimization-level control. While Focal Loss explicitly reshapes gradient contributions, its performance remains highly sensitive to static hyperparameter choices. In contrast, automated search enables adaptive calibration of optimization behavior within the constrained LoRA space, reducing performance volatility across configurations. This suggests that, under extreme imbalance, tuning the training dynamics itself is as critical as modifying the loss formulation.

6 Discussion and Analysis

6.1 Error Analysis and Confusion Patterns

To evaluate qualitative performance, we analyzed the confusion matrix generated from the official test set (Figure 2). The analysis reveals two prominent sources of confusion:

- **Neutral vs. Opinionated Ambiguity:** The model frequently misidentifies *Neutral* reports as *Opinionated* due to high lexical overlap. Distinguishing objective reporting from subtle political bias remains challenging without broader pragmatic cues.
- **Sarcasm Recognition Bottlenecks:** *Sarcastic* instances are often confounded with *Opinionated* and *Neutral* labels. This potentially stems from complex phonetic wordplay in code-mixed Tanglish which current preprocessing cannot explicitly normalize, leading to sub-optimal representation.

Despite these bottlenecks, the *Substantiated* class maintains non-trivial recall under severe data sparsity. This indicates a partial retention of minority-class signals, demonstrating that the optimization regime does not entirely suppress underrepresented categories.

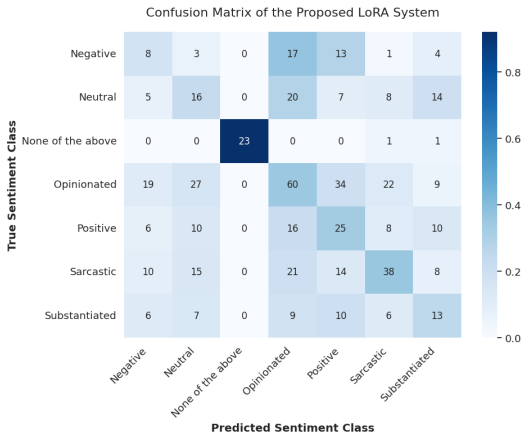

Figure 2: Confusion matrix of the XLM-R + LoRA system.

6.2 Optimization in Bias Mitigation

Compared to standard full fine-tuning, the constrained LoRA parameterization combined with automated hyperparameter search resulted in more balanced confusion patterns. This suggests that identifying a specific hyperparameter configuration empirically correlates with improved minority-class performance, yielding more balanced predictions across sentiment categories under skewed supervision.

7 Conclusion

This paper presented a high-performing pipeline for Political Multiclass Sentiment Analysis of code-mixed Tamil comments. By integrating emoji-aware preprocessing with Low-Rank Adaptation and random oversampling, our system achieved Rank 2 (macro-F1: 0.3763). While this absolute score underscores the inherent complexity of the code-mixed Dravidian linguistic landscape, our results suggest that algorithmically optimizing the adaptation process via Optuna provides a practical complement to static interventions, improving macro-level balance in low-resource, severely imbalanced settings without bespoke loss engineering.

Although our study focuses on code-mixed Tamil political discourse, the observed optimization dynamics may extend to other low-resource multilingual classification settings characterized by heavy class skew. These findings motivate further investigation into the interaction between parameter-efficient adaptation and automated hy-

perparameter search in broader cross-lingual contexts.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

Limitations

While our pipeline achieved a competitive Rank 2, the absolute macro-F1 score of 0.3763 highlights the severe difficulty of the code-mixed Tanglish domain. A primary limitation is the persistent confounding of *Sarcastic*, *Neutral*, and *Opinionated* classes. We hypothesize this stems from complex phonetic wordplay and high lexical overlap, which standard tokenizers and our current preprocessing cannot fully resolve. Furthermore, our methodology relies on integrating existing optimization frameworks rather than proposing novel architectures. Addressing these systemic ambiguities will require exploring bespoke loss functions and architectural designs explicitly tailored for phonetic variations in low-resource settings.

Ethical Considerations

This research strictly utilizes the publicly available dataset provided by the DravidianLangTech@ACL 2026 organizers. We do not attempt to deanonymize users or extract personally identifiable information. The models developed are intended for academic research in computational linguistics and should not be deployed for automated moderation without further fairness and bias evaluations.

AI Usage Acknowledgment

Following the ACL 2026 guidelines, we acknowledge the use of Large Language Models (LLMs) solely for the purposes of English proofreading and LaTeX formatting assistance during the preparation of this manuscript. The authors assume full responsibility for the final content and findings.

Reproducibility

The source code, hyperparameter search scripts, and training configurations are publicly available at: https://github.com/nguyenbbt/ACL2026_DravidianLangTech

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Offensive language identification in low-resourced code-mixed Dravidian languages using Multilingual Transformers. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 104–110.

Edward J Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563.

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*, pages 2980–2988.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, S. Rajalakshmi, and Premjith B. 2021. Named entity recognition for code-mixed Indian languages using Transformer models. In *Proceedings of the 1st Workshop on Speech and Vision for Dravidian Languages*, pages 1–7.

Mani Vegupatti, Kishore Kumar Ponnusamy, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Durairaj Thenmozhi, Prasanna Kumar Kumaresan, and Sathiyaraj Thangasamy. 2026. TamilPoliSent 2026: A Shared Task report on Multiclass Political Sentiment Analysis in Tamil. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.