PolyTicsTamil_Alchemists@DravidianLangTech@ACL 2026: An Augmentation-Driven Focal Ensemble Model for Political Sentiment Analysis in Tamil

Jyoti Kumari[1] **Meclin A Francis**[2] **Vinay Babu Ulli**[3]
Malavika Sreekumar[4] **Joel Johnson**[5]
[1]Department of Linguistics, Banaras Hindu University, Varanasi, India
[2]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, 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 DravidianLangTech@ACL 2026 shared task on Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments. The task requires classifying Tamil political tweets into seven sentiment categories. We address two key challenges, severe class imbalance and semantic overlap between categories, through a three-stage pipeline. First, we balance the training set by augmenting minority classes via back-translation and transformer-based paraphrasing. Second, we fine-tune XLM-RoBERTa-base using a class-weighted Focal Loss (γ=2), which directs learning towards hard, ambiguous samples. Third, we train five models under Stratified 5-Fold Cross-Validation and average their softmax outputs at inference time. On the official test set, the system achieves a Macro-F1 of **0.3539**. The code is publicly available.[1]

1 Introduction

Social media platforms, X (formerly Twitter) in particular, have emerged as primary channels for political discourse. Automatically analysing the sentiment of such content can inform political communication research and public opinion tracking. However, political text is challenging for sentiment analysis due to rhetorical devices, sarcasm, and culturally coded references.

These difficulties are compounded for Tamil, which is morphologically rich and agglutinative. Political commentary on Tamil social media further exhibits pervasive code-mixing with English, non-standard orthography, and dialectal variation.

The DravidianLangTech@ACL 2026 shared task (Vegupatti et al., 2026) on *Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments* formalises this as a seven-way classification task with labels: *Negative*, *Neutral*, *None of the above*, *Opinionated*, *Positive*, *Sarcastic*, and *Substantiated*. Two aspects make this task particularly difficult: the label distribution is heavily skewed, and certain label pairs such as *Opinionated* versus *Sarcastic* share overlapping surface features.

We address these challenges through three stages: (1) synthetic augmentation of underrepresented classes using back-translation and paraphrasing until all classes reach equal frequency; (2) fine-tuning XLM-RoBERTa-base (Conneau et al., 2020) with class-weighted Focal Loss (Lin et al., 2017); and (3) ensembling five models trained under Stratified 5-Fold Cross-Validation via probability averaging.

2 Related Work

The DravidianLangTech workshop series has highlighted the difficulties of processing Tamil social media text (Chakravarthi et al., 2020). The preceding DravidianLangTech@NAACL 2025 iteration introduced this political sentiment task with the same seven-class label set (Chakravarthi et al., 2025), where overall Macro-F1 scores remained modest despite 25 participating teams. The top-ranked Synapse team combined IndicBERTv2-MLM features with TF-IDF and back-translation (Kp et al., 2025). Roy et al. (2025) fine-tuned L3Cube-Tamil-BERT, while Shanmugavadivel et al. (2025) paired Random Forest with feedforward networks. S et al. (2025) employed SGD with incremental learning, and K et al. (2025) embedded tweets with LaBSE and classified with SVMs.

Beyond this shared-task series, processing low-resource Indian languages like Tamil has seen growing interest, with researchers frequently grappling with code-mixing and morphological complexity. Furthermore, handling multi-class imbalance in text classification is a widely recognized challenge in the broader NLP literature. Researchers frequently employ oversampling, cost-sensitive learn-

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

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 326–330

ing, and advanced objective functions to prevent majority class dominance.

A common limitation is that class imbalance is typically addressed at only one level. Our work departs by enforcing class parity through augmentation, pairing it with Focal Loss for hard-example mining, and aggregating predictions across five folds to mitigate overfitting.

3 Methodology

We frame the task as a multi-class classification problem over seven political sentiment categories. Our system consists of three stages: data augmentation to address class imbalance, fine-tuning of a pre-trained transformer encoder, and cross-validation ensembling with focal loss.

3.1 Data Augmentation

Let N_c denote the number of training samples in class c, and $N_{\max} = \max_c N_c$. For every class where $N_c < N_{\max}$, we generate $N_{\max} - N_c$ additional samples using two complementary techniques. Back-translation translates each Tamil tweet to English and back to Tamil using Google Translate, producing structural paraphrases. Transformer-based paraphrasing uses IndicBART (Dabre et al., 2022) to generate lexical and syntactic variants. We alternate between the two techniques until all classes reach $N_{\max}$ samples, yielding a uniform training distribution.

3.2 Model Architecture and Loss Function

We use XLM-RoBERTa-base (Conneau et al., 2020) as our encoder. The `[CLS]` token representation $\mathbf{z} \in \mathbb{R}^{768}$ is passed through a dropout layer (p=0.4) and a linear projection to produce logits $\mathbf{o} \in \mathbb{R}^7$.

We train with class-weighted Focal Loss (Lin et al., 2017), which introduces a modulating factor $(1-p_t)^\gamma$ that suppresses the loss for well-classified examples:

$$\mathcal{L}_{\text{FL}} = -\alpha_c \, (1 - p_t)^\gamma \, \log(p_t) \qquad (1)$$

where p_t is the predicted probability for the ground-truth class, γ=2 is the focusing parameter, and $\alpha_c = N_{\text{total}}/(C \cdot N_c)$ is the per-class weight.

3.3 Stratified k-Fold Ensemble

We employ Stratified 5-Fold Cross-Validation to reduce variance from any single split. Five independent XLM-R models are trained, each using four folds for training and one for validation. At inference, all five models produce softmax vectors and the final prediction is obtained by averaging:

$$\hat{y} = \arg\max_j \; \frac{1}{K} \sum_{i=1}^{K} \text{softmax}(\text{Model}_i(x))_j \qquad (2)$$

where K=5. This soft-voting strategy preserves confidence information, allowing highly certain models to exert proportionally greater influence.

Category	Original	After Aug.
Opinionated	1,361	1,361
Sarcastic	790	1,361
Neutral	637	1,361
Positive	575	1,361
Substantiated	412	1,361
Negative	406	1,361
None of the above	171	1,361
Total	4,352	9,527

Table 1: Class distribution before and after augmentation.

Hyperparameter	Value
Pre-trained model	`xlm-roberta-base`
Optimizer	AdamW
Learning rate	1×10^{-5}
LR scheduler	Linear with warm-up
Batch size	16
Epochs per fold	5
Max sequence length	128
Dropout rate	0.4
Focal Loss γ	2.0
Number of folds (K)	5

Table 2: Hyperparameter configuration.

4 Experimental Setup

4.1 Dataset

Table 1 reports the original class distribution. After augmentation, minority classes are expanded to match the majority class count ($N_{\max}$=1,361). The test set contains 544 unlabeled samples.

4.2 Implementation Details

Our pipeline was implemented in PyTorch using the Hugging Face `transformers` library (Wolf et al., 2020) on a single NVIDIA GPU. Table 2 summarises the hyperparameters. For each fold, we saved the checkpoint with the highest validation Macro-F1.

Configuration	Val F1
XLM-R, no aug., CE loss	0.2200
XLM-R, aug., Focal Loss	0.4268
5-Fold Ensemble, aug., FL	**0.4050**

Table 3: Ablation results. CE = Cross-Entropy, FL = Focal Loss. The single-split augmented model reports best validation F1; the ensemble reports mean CV F1.

Fold	Val Macro-F1
1	0.4158
2	0.3900
3	0.4038
4	0.4083
5	0.4069
Mean	**0.4050**

Table 4: Per-fold validation Macro-F1 scores.

Metric	Macro	Weighted
Accuracy	0.3474	
Precision	0.3493	0.3304
Recall	0.3722	0.3474
F1-Score	**0.3539**	0.3328

Table 5: Official test set results.

Class	P	R	F1
Negative	0.27	0.29	0.28
Neutral	0.52	0.12	0.20
None of the above	0.94	0.88	0.91
Opinionated	0.34	0.45	0.39
Positive	0.31	0.48	0.38
Sarcastic	0.54	0.48	0.51
Substantiated	0.37	0.30	0.33
Macro Avg	0.47	0.43	**0.43**

Table 6: Class-wise precision (P), recall (R), and F1 for the best single-split model on the validation set.

5 Results and Analysis

5.1 Validation Results

Table 3 presents an ablation comparing three configurations on the validation set. The baseline XLM-R model trained with standard cross-entropy on the original imbalanced data achieves a Macro-F1 of 0.2200. Adding augmentation and Focal Loss improves this to 0.4268, confirming that addressing imbalance at both data and loss-function levels yields complementary gains. The 5-Fold ensemble produces stable per-fold F1 scores (Table 4), with a mean cross-validation Macro-F1 of 0.4050.

Interestingly, the ensemble's mean CV Macro-F1 (0.4050) is lower than the best single-split model (0.4268). This suggests that our simple soft-voting strategy may occasionally over-average predictions, diluting highly confident correct outputs from individual folds.

5.2 Official Test Results

Table 5 reports the official test set results. The system achieves a Macro-F1 of **0.3539** with an accuracy of 0.3474. The drop from validation Macro-F1 (0.4268 single-split; 0.4050 mean CV) to test Macro-F1 (0.3539) indicates that the augmented training distribution does not fully capture the diversity of the test set. The gap between Macro-F1 (0.3539) and weighted F1 (0.3328) further suggests that the system performs relatively better on minority classes than on majority classes, consistent with the augmentation and Focal Loss strategy that explicitly upweights underrepresented categories.

5.3 Class-wise Analysis

Table 6 reports per-class performance from the best single-split model on the validation set. The system performs strongly on *None of the above* (F1 = 0.91), which is the most lexically distinct category. Performance on *Sarcastic* (0.51) is moderate, while *Neutral* (0.20) and *Negative* (0.28) remain challenging due to overlapping surface features with other categories.

5.4 Error Analysis

Three dominant error patterns emerge from manual inspection. First, **sarcasm–opinion confusion**: sarcastic tweets often adopt the surface structure of genuine opinions, differing only in pragmatic intent, leading to frequent misclassification between *Sarcastic* and *Opinionated*. Second, **low recall for Neutral**: the model struggles to identify neutral comments, as they lack the strong lexical signals that characterise opinionated or sarcastic text, resulting in many neutral samples being absorbed into adjacent categories. Third, **augmentation noise**: we did not formally assess the semantic validity of every generated sample. Manual inspection indicates that while back-translation and paraphrasing successfully increase class frequencies, they occasionally introduce translationese artifacts or alter the pragmatic meaning of tweets due to the uneven quality of machine translation for Tamil. This noise strongly contributes to the test-set generalization gap.

6 Conclusion

We presented our system for the DravidianLangTech@ACL 2026 shared task on Political Multiclass Sentiment Analysis of Tamil Twitter Comments (Vegupatti et al., 2026). Our three-stage pipeline addresses class imbalance through back-translation and paraphrasing-based augmentation, class-weighted Focal Loss (γ=2), and a Stratified 5-Fold ensemble. The ablation confirms that each stage contributes incremental gains, with the augmentation and Focal Loss combination yielding the largest improvement (from 0.2200 to 0.4268 validation Macro-F1). On the official test set, the system achieves a Macro-F1 of 0.3539. The gap between validation and test performance suggests that synthetic augmentation, while effective for balancing class frequencies, does not fully substitute for natural data diversity. Future work will explore incorporating discourse-level features to address sarcasm–opinion confusion, learned ensemble weights via a meta-classifier, and larger Indic-specific pre-trained models such as MuRIL as the backbone encoder.

Limitations

Our work has several limitations. First, the augmentation strategy relies on back-translation through English via Google Translate and paraphrasing with IndicBART, both of which may introduce translationese artefacts or semantic drift that do not reflect the natural distribution of Tamil political discourse; the gap between validation Macro-F1 (0.4050) and test Macro-F1 (0.3539) supports this concern. Second, we use only XLM-RoBERTa-base as our encoder and do not compare against Indic-specialised models such as MuRIL or IndicBERT, which may capture Tamil morphological patterns more effectively. Third, the seven sentiment categories exhibit inherent semantic overlap—particularly between *Sarcastic* and *Opinionated* that surface-level token representations struggle to disambiguate, yet we do not incorporate any discourse-level, pragmatic, or contextual features beyond the individual tweet. Fourth, the Focal Loss hyperparameter γ=2 and the dropout rate of 0.4 were selected based on common defaults rather than through systematic tuning, and may not be optimal for this label distribution. Fifth, our error analysis is qualitative and based on manual inspection of a small sample; we do not provide a formal confusion matrix on the test set or a quantitative breakdown of error types. Finally, the system is evaluated on a single dataset from one shared task; its ability to generalise to other Tamil sentiment benchmarks, different time periods, or other Dravidian languages remains untested.

References

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A sentiment analysis dataset for code-mixed Malayalam-English. 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 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Arunaggiri Pandian Karunanidhi, and Rohan R. 2025. Overview on political multiclass sentiment analysis of Tamil X (Twitter) comments: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 746–753, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. 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.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Nithish Ariyha K, Eshwanth Karti T R, Yeshwanth Balaji A P, Vikash J, and Sachin Kumar S. 2025. Wictory@DravidianLangTech 2025: Political sentiment analysis of Tamil X(Twitter) comments using LaBSE and SVM. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 706–710, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Suriya Kp, Durai Singh K, Vishal A S, Kishor S, and Sachin Kumar S. 2025. Synapse@DravidianLangTech 2025: Multiclass political sentiment analysis in Tamil X

(Twitter) comments: Leveraging feature fusion of IndicBERTv2 and lexical representations. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 716–720, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Billodal Roy, Pranav Gupta, Souvik Bhattacharyya, and Niranjan Kumar M. 2025. LexiLogic@DravidianLangTech 2025: Multimodal hate speech detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 552–556, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Kalaivani K S, Sanjay R, Thissyakkanna S M, and Nirenjhanram S K. 2025. KSK@DravidianLangTech 2025: Political multiclass sentiment analysis of Tamil X (Twitter) comments using incremental learning. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 221–225, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Sanjai R, Mohammed Sameer, and Motheeswaran K. 2025. Beyond_Tech@DravidianLangTech 2025: Political multiclass sentiment analysis using machine learning and neural network. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 139–143, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

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.