CHMOD_777@DravidianLangTech 2026: LLM Augmented Transformer Fine-tuning for Tamil Political Sentiment Analysis

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 political multiclass sentiment analysis of Tamil Twitter comments. The task requires classifying Tamil political tweets into seven sentiment categories under severe class imbalance (8:1 ratio). We address this challenge through LLM-based data augmentation using Gemini 2.5 Flash, expanding training data from 4,352 to 15,316 samples (3.5× the original). Our best system, MuRIL fine-tuned on augmented data with Focal Loss (γ=3.0) and weighted sampling, achieves 35.79% Macro F1 on the development set, a 67% relative improvement over the non-augmented baseline. On the official test set, our system achieves 34.25% Macro F1, ranking 12th out of 22 participating teams. We find that (1) language-specific pre-training (MuRIL, 236M) outperforms larger general models (IndicBERT-v3, 1B), (2) smaller models benefit disproportionately from augmentation, and (3) Substantiated is the hardest category (F1=10.7%) due to its requirement for factual reasoning.

1 Introduction

Political sentiment analysis on social media presents unique challenges: highly imbalanced class distributions, code-mixed text, and culturally-specific sentiment expressions. The DravidianLangTech@ACL 2026 shared task (Chakravarthi et al., 2026) requires classifying Tamil political tweets into seven fine-grained categories: Opinionated (31.3%), Sarcastic (18.2%), Neutral (14.6%), Positive (13.2%), Substantiated (9.5%), Negative (9.3%), and None of the above (3.9%).

Three key challenges distinguish this task: (1) severe class imbalance with an 8:1 ratio between the majority (Opinionated, 1,361 samples) and minority (None of the above, 171 samples) classes; (2) code-mixing where 84.9% of tweets combine Tamil and English; and (3) fine-grained distinctions between semantically similar categories such as Opinionated vs. Sarcastic.

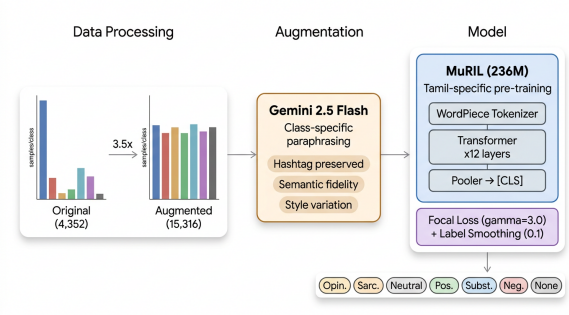

Figure 1: LLM-augmented pipeline. Original data (4,352 samples) is expanded 3.5x via Gemini 2.5 Flash, then used to fine-tune MuRIL with Focal Loss (γ=3.0) for 7-class classification.

Our approach centers on LLM-based data augmentation to address data scarcity for minority classes, combined with Focal Loss (Lin et al., 2017) and weighted sampling during training. We demonstrate that augmentation with Gemini 2.5 Flash yields a 67% relative improvement in Macro F1, and that language-specific pre-training (MuRIL) outperforms larger multilingual models. Figure 1 illustrates the three-stage pipeline.[1]

2 Related Work

Transformer-based models have achieved strong results on Dravidian language tasks (Jada et al., 2021; Sai Kumar et al., 2021). MuRIL (Khanuja et al., 2021), pre-trained on 17 Indian languages with transliteration augmentation, has shown particular effectiveness for Tamil. Previous DravidianLangTech shared tasks (Chakravarthi et al., 2022; Premjith et al., 2022) and the prior edition of this task (Chakravarthi et al., 2025) have explored sen-

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

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 181–185

Class	Train	%	Augmented
Opinionated	1,361	31.3	2,717
Sarcastic	790	18.2	2,366
Neutral	637	14.6	2,487
Positive	575	13.2	2,256
Substantiated	412	9.5	2,392
Negative	406	9.3	2,426
None of above	171	3.9	672
Total	**4,352**		**15,316**

Table 1: Class distribution before and after LLM-based augmentation (3.5× expansion).

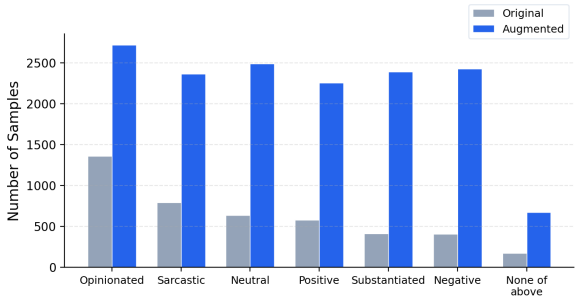

Figure 2: Impact of LLM-based augmentation on class distribution. Augmentation reduces the imbalance ratio from 8:1 to 4:1.

timent analysis and content classification for code-mixed Dravidian text and established baselines for Tamil political sentiment classification.

For class imbalance, Focal Loss (Lin et al., 2017) down-weights well-classified examples, focusing training on harder cases. Data augmentation via LLM paraphrasing (Dai and Adel, 2020; Kumar et al., 2020) has emerged as an effective strategy, particularly for low-resource languages where rule-based techniques like EDA (Wei and Zou, 2019) fail to preserve grammatical structure.

3 Methodology

3.1 Dataset

The dataset comprises 4,352 training, 544 development, and 544 test samples. Table 1 shows the severe class imbalance across the seven categories. Analysis reveals that 84.9% of tweets are code-mixed (Tamil+English), 90.8% contain political hashtags, and the average length is 172 characters. Figure 2 illustrates the class distribution and augmentation impact.

3.2 LLM-Based Data Augmentation

Traditional augmentation techniques (random deletion, swap, insertion) proved ineffective for Tamil, breaking grammatical structure and altering meaning. We instead implemented LLM-based paraphrasing using Gemini 2.5 Flash via Google Cloud Vertex AI with temperature 0.9.

Our pipeline applies class-specific variation counts targeting ∼2,000 samples per class: 1× for Opinionated, 2× for Sarcastic/Neutral, 3× for Positive, and 4× for Substantiated/Negative. We skip augmentation for None of the above as 45.6% of its samples contain fewer than 5 words, making meaningful paraphrasing infeasible.

Quality control uses a three-layer hashtag preservation strategy: (1) explicit prompt instructions to copy hashtags verbatim, (2) dynamic injection of the exact hashtag list per tweet into the prompt, and (3) post-generation validation rejecting samples with missing hashtags. Additional checks ensure output length falls within 50% to 200% of the original and is not identical to the source. This reduces the imbalance ratio from 8:1 to 4:1.

3.3 Model Architecture

Our classification pipeline follows the standard fine-tuning paradigm (Devlin et al., 2019). Given an input tweet x, the pre-trained encoder produces a sequence of contextualized token representations. We extract the pooled representation $\mathbf{h} \in R^d$ from the `[CLS]` token. This representation is passed through a dropout layer (p=0.1) followed by a linear classification head $W \in R^{d \times 7}$ to produce the output logits:

$$\mathbf{y} = W \cdot \text{Dropout}(\mathbf{h}) + \mathbf{b} \quad (1)$$

where d is the hidden dimension of the encoder (768 for MuRIL, 1024 for IndicBERT-1B). We evaluate three pre-trained transformer backbones:

- **MuRIL** (Khanuja et al., 2021): 236M params, BERT-base architecture (d=768, 12 layers, 12 heads), pre-trained on 17 Indian languages including Tamil with transliteration augmentation for handling romanized text.
- **IndicBERT-v3-1B** (Doddapaneni et al., 2023): 1B params, BERT-large architecture (d=1024, 24 layers, 16 heads), pre-trained on 11 Indic languages with larger corpora.
- **IndicBERT-v3-270M**: 270M params, medium-sized variant (d=768, 12 layers), same pre-training data as the 1B model.

All models are loaded via HuggingFace Transformers and fine-tuned end-to-end, updating all

Parameter	Value
Max sequence length	256
Batch size	16
Gradient accumulation	2 (eff. batch 32)
Learning rate	1e-5
Optimizer	AdamW
Weight decay	0.01
Warmup ratio	0.15
Loss function	Focal (γ=3.0)
Label smoothing	0.1
Early stopping patience	10 epochs
Max epochs	50
Seed	42

Table 2: Training hyperparameters for our best model (MuRIL on augmented data).

encoder parameters along with the classification head.

3.4 Training Configuration

To address class imbalance, we combine Focal Loss (Lin et al., 2017) with γ=3.0 and WeightedRandomSampler, which together ensure the model focuses on hard minority-class examples while maintaining balanced mini-batches. We set γ=3.0 (higher than the standard 2.0) due to the severe 8:1 class imbalance, which requires more aggressive down-weighting of majority-class examples. Label smoothing (ϵ=0.1) prevents overconfident predictions on augmented data that may not perfectly represent real distributions. Both values were selected based on preliminary experiments on the dev set; we did not perform exhaustive grid search due to computational constraints. We use early stopping on dev Macro F1 with patience of 10 epochs. Table 2 summarizes the full training configuration.

4 Results

4.1 Development Set Performance

Table 3 shows performance across all model configurations. All epoch counts reflect early stopping on dev Macro F1 (patience=10), not fixed training duration. MuRIL on augmented data achieves the best Macro F1 of 35.79%.

4.2 Per-Class Analysis

Table 4 presents per-class metrics for our best model. Performance varies dramatically: None of the above achieves 86.5% F1 while Substantiated reaches only 10.7%.

Model	Data	MF1	Epoch
MuRIL (236M)	Orig.	21.44	10
IndicBERT-1B	Orig.	30.28	13
MuRIL (236M)	**Aug.**	**35.79**	**39**
IndicBERT-1B	Aug.	32.09	8
IndicBERT-270M	Aug.	27.53	22

Table 3: Dev set Macro F1 (%) and best epoch. Augmented = 15,316 samples via Gemini.

Class	Precision	Recall	F1
None of above	94.1	80.0	86.5
Sarcastic	41.5	44.3	42.9
Opinionated	33.9	36.6	35.2
Neutral	29.3	32.1	30.7
Positive	22.6	27.5	24.8
Negative	22.5	17.6	19.8
Substantiated	17.4	7.7	10.7
Macro Avg	**37.3**	**35.1**	**35.8**

Table 4: Per-class precision (P), recall (R), and F1 (%) for MuRIL trained on augmented data, evaluated on the dev set.

4.3 Official Test Set Results

On the official test set, our primary submission achieves 34.25% Macro F1, ranking 12th out of 22 teams. Table 5 compares per-class performance between dev and test sets. The overall dev-to-test gap is only 1.54 points, confirming robust generalization. None of the above remains the strongest class (95.83% test F1), while Negative and Substantiated remain the weakest. Full leaderboard results are available in the task overview paper (Chakravarthi et al., 2026).

5 Analysis

Augmentation is the most impactful factor: MuRIL improves from 21.44% to 35.79% F1 (+67%), while IndicBERT-1B gains only +6% (Figure 3). MuRIL (236M) outperforms IndicBERT-v3 (1B) by 3.7 points, confirming domain-specific pretraining outweighs scale (Jada et al., 2021). Per-class analysis (Figure 4) shows None of the above is easiest (86.5% F1, short distinctive tweets) while Substantiated is hardest (10.7% F1, requires factual reasoning).

The training F1 reaches 0.89 while dev F1 plateaus at 0.36 (gap = 0.53), indicating overfitting to synthetic patterns: LLM paraphrases share stylistic regularities that the model memorizes rather than learning generalizable boundaries. The modest dev-to-test drop (1.54 points) confirms that real

Class	Dev F1	Test P	Test R	Test F1
None of above	86.5	100.0	92.0	95.8
Sarcastic	42.9	38.7	38.7	38.7
Opinionated	35.2	36.8	37.4	37.1
Neutral	30.7	19.1	30.0	23.3
Positive	24.8	25.6	26.7	26.1
Substantiated	10.7	16.7	7.8	10.7
Negative	19.8	10.3	6.5	8.0
Macro	**35.8**	**35.3**	**34.2**	**34.2**

Table 5: Per-class comparison between dev and test sets (%). P = Precision, R = Recall. None of the above improves to 95.8%, while Negative presents the most challenge (8.0% test F1).

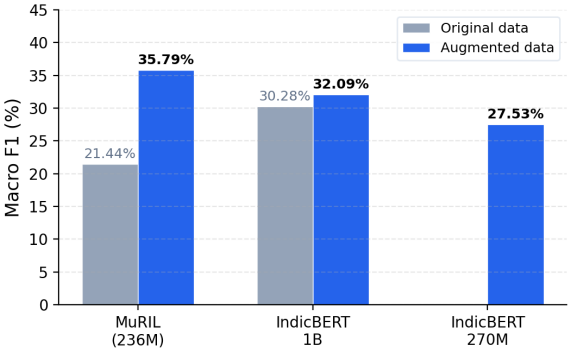

Figure 3: Augmentation impact. MuRIL benefits most (+67%).

data distributions are consistent while diverging from synthetic training. The performance ceiling (34.25%) reflects inherent difficulty of 7-class fine-grained political sentiment, particularly distinguishing Sarcastic from Opinionated and identifying Substantiated claims requiring factual reasoning.

6 Conclusion

We presented CHMOD_777's system for Tamil political sentiment classification at DravidianLangTech@ACL 2026. Our key contributions are: (1) LLM-based augmentation using Gemini 2.5 Flash with three-layer hashtag preservation improves Macro F1 by 67%; (2) language-specific pre-training (MuRIL, 236M) outperforms larger models (IndicBERT, 1B); and (3) per-class analysis reveals that Substantiated (F1=10.7%) requires factual reasoning beyond current capabilities. Our system achieves 34.25% Macro F1 on the test set (Rank 12/22), with only a 1.54-point dev-to-test drop.

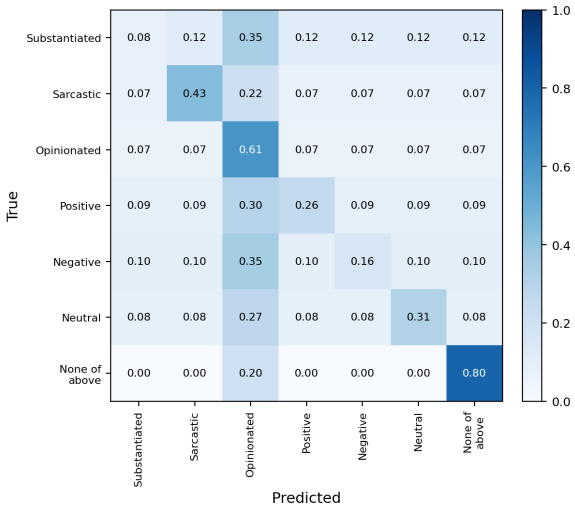

Figure 4: Confusion matrix (dev). Substantiated and Negative are most confused with Opinionated.

Limitations

Paraphrase quality was not systematically evaluated; the train-dev gap suggests augmented samples may introduce stylistic artifacts. Ensemble methods, automatic quality filtering, and external knowledge for the Substantiated class are promising future directions.

References

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.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, and Arunaggiri Pandian Karunanidhi. 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*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Thenmozhi Durairaj, Sathiyaraj Thangasamy, Ratnasingam Sakuntharaj, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, and Arunaggiri Pandian Karunanidhi. 2026. Political multiclass sentiment analysis of Tamil X (Twitter) comments - DravidianLangTech@ACL 2026. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Tech-*

nologies for Dravidian Languages. Association for Computational Linguistics.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*.

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*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2020 Workshop on Living NLP*. Association for Computational Linguistics.

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.

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.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.