ByteBuilders@DravidianLangTech 2026: Transformer-Based Weighted Ensemble for Political Multiclass Sentiment Analysis of Tamil X (Twitter) Comments

Mitharshana T V[1] **and Shanthi S**[1] **and Lavana V**[1] **and Kaviya Varma R**[1]
[1]Department of Computer Science and Engineering
Kongu Engineering College, Perundurai, India
mitharshanatv.22cse@kongu.edu, shanthis@kongu.ac.in,
lavanav.22cse@kongu.edu, kaviyavarmar.22cse@kongu.edu

Abstract

Our work is submitted as part of the DravidianLangTech 2026 shared task on Tamil political sentiment analysis. Tamil political sentiment analysis still focuses more difficulties due to linguistic diversity, informal social media expressions, sarcasm, and severe class imbalance in low-resource languages. Our paper presents a transformer based weighted ensemble approach for 7 class political sentiment classification of Tamil X (Twitter) comments. The proposed system combines XLM-RoBERTa and IndicBERT using weighted logit averaging to leverage both multilingual semantic understanding and Indic-specific linguistic representations.

To address class imbalance, random oversampling, balanced class weights, and focal loss with $\gamma = 2.0$ are employed. Additionally, stratified 10-fold cross validation is used to improve robustness and it reduces overfitting. Experimental results show that the proposed ensemble model achieves an accuracy of 36.8% and a weighted F1-score of 34% on the development dataset. The results demonstrate the difficulties involved in fine grained Tamil political sentiment analysis, yet the results of ensemble-based transformer models can provide modest improvements in low-resource Tamil NLP tasks.

Index Terms— Tamil Sentiment Analysis, Transformer Models, Ensemble Learning, XLM-RoBERTa, IndicBERT, Class Imbalance, Focal Loss

1 Introduction

Sentiment analysis or opinion mining, is a fundamental task in Natural Language Processing (NLP) that focuses on identifying and classifying opinions, emotions, and attitudes expressed in textual data. With the rapid growth of social media platforms such as X (Twitter), where users generate enormous volume of data containing their opinions regarding numerous of things, including politics. Analyzing of these kinds of data provides valuable insights for governments, policymakers, and researchers to understand public sentiment, monitor political trends, and support decision-making processes.

Recent advancements in transformer based architectures have significantly improved the performance of NLP systems more effectively than traditional methods. The introduction of the Transformer architecture enabled parallel computation and improved contextual representation learning through self-attention mechanisms (Vaswani et al., 2017). In order to enhance this, BERT introduced bidirectional contextual understanding, improving performance across a wide range of NLP tasks (Devlin et al., 2019). Multilingual models such as XLM-RoBERTa further extended these capabilities to multiple languages, making them particularly effective for low-resource language applications (Conneau et al., 2020).

Even with more advancements, Sentiment analysis of Tamil language remains challenging due to limited annotated data and linguistic complexity. Tamil social media text often contains informal expressions, dialectal variations, code-mixed content, and implicit sentiment, which complicate classification. Political discourse additionally includes sarcasm and rhetorical expressions that are difficult for computational models to interpret accurately (Kakwani et al., 2020). Class imbalance is another major key issue where certain sentiment categories are not presented correctly, which leads to biased predictions and reduces model performance.

To tackle these challenges, this work proposes

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 163–168

a transformer-based weighted ensemble approach combining XLM-RoBERTa and IndicBERT for multiclass sentiment classification of Tamil political comments from the DravidianLangTech 2026 shared task. The proposed framework integrates multilingual semantic representations with Indic-specific linguistic features using weighted logit averaging and stratified 10-fold cross-validation. Furthermore, focal loss, balanced class weights, and oversampling techniques are employed to improve robustness against class imbalance.

The main contributions of this work are summarized as follows:

- A weighted ensemble framework combining XLM-RoBERTa and IndicBERT for Tamil political sentiment classification.
- Integration of focal loss, balanced class weights, and oversampling for handling severe class imbalance.
- Use of stratified 10-fold cross-validation to improve robustness and generalization.
- Comparative evaluation of individual transformer models and the proposed ensemble architecture.

2 Literature Survey

In recent years, there has been tremendous improvement in NLP, primarily by transformer-based architectures, which utilize self-attention mechanisms to model contextual relationships within the text. The Transformer model proposed by Vaswani et al. marked a significant shift from traditional sequential models by enabling parallel computation and capturing long-range dependencies effectively (Vaswani et al., 2017). BERT introduced bidirectional pretraining, allowing models to learn contextual representations from both left and right contexts simultaneously, thereby achieving substantial improvements across multiple NLP tasks (Devlin et al., 2019).

Multilingual transformer models can be achieved by using XLM-RoBERTa model which was developed and trained on large-scale multilingual corpora, demonstrating strong cross-lingual transfer performance (Conneau et al., 2020). Such models are particularly useful for low-resource languages where annotated data is limited. For Indian languages, IndicBERT was introduced to better capture linguistic characteristics such as rich morphology and code-mixed text, which are common in languages like Tamil (Kakwani et al., 2020).

Before the widespread of transformer models, traditional machine learning and deep learning techniques such as Support Vector Machines (SVMs), Logistic regression, and Long Short-Term Memory (LSTM) networks were commonly used for classifying sentiment analysis (Behera et al., 2021). Although LSTM-based models were effective in capturing sequential dependencies in text, their sequential nature limited scalability and contextual understanding compared to transformer-based architectures (He and Garcia, 2009).

Ensemble learning has emerged as an effective strategy for improving model performance by combining predictions from multiple models. This technique ensures the reduction of variance in predictions as well as better generalization by utilizing the advancement of each model(Dietterich, 2000). Handling class imbalance remains another critical challenge in multiclass sentiment classification. Techniques such as focal loss have been proposed to address this issue by emphasizing hard-to-classify samples and reducing the dominance of majority classes during training (Lin et al., 2017).

3 Dataset Description

The dataset used in this work was provided as part of Dravidian LangTech 2026 shared task on Tamil political sentiment analysis. the dataset consists of Tamil political comments collected from the X (Twitter) social media platform. The objective of the task is to classify each comment into one of seven sentiment categories: *Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral,* and *None of the Above* (Behera et al., 2021).

The dataset contains highly diverse social media text with informal writing styles, code mixed expressions, sarcasm, dialectal variations, and implicit sentiment patterns. These characteristics make fine-grained sentiment classification particularly challenging in low-resource Tamil NLP settings (Kohavi, 1995).

Table 1 presents the distribution of sentiment classes in the dataset. The dataset exhibits significant class imbalance, where certain classes such as *Opinionated* and *Sarcastic* contain substantially more samples than minority classes such as *Negative* and *None of the Above*. This class imbalance can bias transformer models towards dominant classes and reduce minority-class prediction

Table 1: Dataset distribution across sentiment classes

Class	Samples
Substantiated	412
Sarcastic	790
Opinionated	1361
Positive	575
Negative	406
Neutral	637
None of the Above	171

performance.

To address this issue, random oversampling was applied to minority classes during training. In addition, balanced class weights and focal loss were employed to improve learning for underrepresented sentiment categories.

4 System Architecture and Implementation

The proposed system employs a transformer-based ensemble architecture combining XLM-RoBERTa and IndicBERT for Tamil political sentiment classification. The framework integrates multilingual semantic representations with Indic-specific linguistic features using weighted logit averaging and stratified 10-fold cross-validation.

Figure 1 illustrates the complete workflow of the proposed system. The architecture combines multilingual contextual representations from XLM-RoBERTa with Indic-specific linguistic features extracted using IndicBERT. The ensemble framework integrates outputs from both models using weighted averaging to improve robustness and generalization (Dietterich, 2000).

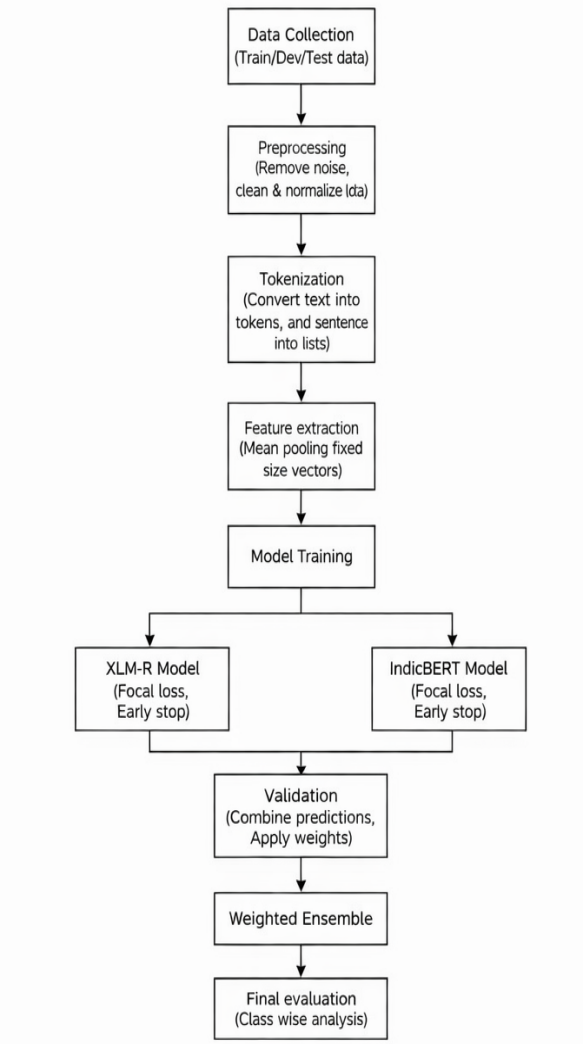

Figure 1: Overall workflow of the proposed transformer-based ensemble architecture. The input Tamil comments are processed through XLM-RoBERTa and IndicBERT models independently. Mean pooling, focal loss, oversampling, and stratified K-fold validation are applied before combining predictions using weighted logit averaging.

4.1 Base Models

Two pretrained transformer models are utilized in this work:

- XLM-RoBERTa-Base, developed by Meta AI and pretrained on large-scale multilingual corpora.
- IndicBERT, developed by AI4Bharat at IIT Madras and optimized for various Indic languages including Tamil.

The proposed system employs these models due to their complementary strengths in multilingual and Indic language representation learning. XLM-RoBERTa captures rich multilingual semantic relationships across languages (Conneau et al., 2020) (Conneau and Lample, 2019), while IndicBERT effectively models linguistic and morphological characteristics specific to Indian languages such as Tamil (Kakwani et al., 2020).

Every input comment is tokenized using the corresponding tokenizer associated with the pretrained models. A maximum sequance of 256 tokens is used, where shorter sequence are padded and longer sequence are truncated. Attention masks are generated to distinguish between valid tokens and padding tokens during model processing.

The tokenized inputs are passed through the transformer encoders to obtain contextualized token embeddings. Instead of relying only on the special CLS token representation, mean pooling is applied over valid token embeddings to obtain a more stable sentence-level representation. This approach is particularly effective for noisy and informal social media text. The pooled representation is then passed through a dropout layer with a probability of 0.3, followed by a fully connected linear layer that produces logits corresponding to the seven sentiment classes (Vegupatti et al., 2026).

4.2 Stratified 10-Fold Cross-Validation

A 10-fold cross-validation technique is employed to improve model reliability and reduce overfitting. This technique preserves class distribution across folds while enabling robust evaluation on multiple train-validation splits (Jurafsky and Martin, 2023).

For each fold, 9 subsets are used for training and 1 subset is used for validation. Validation accuracy is monitored after each epoch, and the best-performing model state is saved. Early stopping with a patience of three epochs is applied to prevent overfitting.

This process produces ten trained XLM-RoBERTa models and ten trained IndicBERT models. Final predictions are generated by averaging the logits obtained from all folds during inference.

4.3 Model Ensembling and Class Imbalance Handling

To enhance robustness and predictive performance, a weighted ensemble strategy is employed to combine the outputs of XLM-RoBERTa and IndicBert.

$$L = w_1 L_{XLMR} + w_2 L_{IndicBERT} \quad (1)$$

where L_{XLMR} and $L_{IndicBERT}$ denote the logits produced by XLM-RoBERTa and IndicBERT respectively, while w_1 and w_2 represent the ensemble weights.

Development set accuracy determines the empirical weights. The final weights assigned were 0.499 for XLM-RoBERTa and 0.501 for IndicBERT.

The dataset used in this work exhibits significant class imbalance. To overcome this, Data-level imbalance and algorithm-level imbalance strategies are employed.

At the data level, random oversampling is applied to increase the representation of minority classes. At the algorithm level, balanced class weights and focal loss with $\gamma = 2.0$ are used during training. Unlike standard cross-entropy loss, focal loss emphasizes hard-to-classify samples, improving learning for underrepresented classes.

4.4 Training Configuration

The models were trained using the AdamW optimizer with a learning rate of 1.5×10^{-5} and a batch size of 16. The maximum number of tokens was fixed to 256 tokens. The models were trained upto 15 epochs, with an early stopping patience of 3 epochs. A dropout probability of 0.3 was applied before the classification layer. Gradient clipping with a maximum norm of 1.0 and a linear warmup scheduler were used to stabilize training.

5 Results and Discussion

5.1 Individual Model and Ensemble Performance

Table 2: Performance comparison of individual transformer models and the proposed weighted ensemble model on the development set.

Model	Accuracy
XLM-RoBERTa	0.3438
IndicBERT	0.3456
Proposed Weighted Ensemble	0.3676

Table 2 presents the performance comparison between the individual transformer models and the proposed weighted ensemble approach. Both XLM-RoBERTa and IndicBERT achieve comparable development accuracies, while the ensemble model demonstrates improved performance by combining multilingual semantic understanding with Indic-specific linguistic representations (Vaswani et al., 2017).

The proposed transformer-based weighted ensemble model achieved an overall development-set accuracy of 36.8% and a weighted F1-score of 34%. Although the improvements are moderate, the ensemble approach provides better robustness and generalization compared to individual models.

5.2 Classification Performance

Table 3 presents the detailed class-wise evaluation metrics of the proposed ensemble model. The model performs relatively well on the Opinionated and Sarcastic classes, while the performance remains lower for Negative and Neutral classes due

Table 3: Class-wise precision, recall, F1-score, and support for the proposed ensemble model.

Class	Precision	Recall	F1-score	Support
Substantiated	0.22	0.17	0.19	52
Sarcastic	0.51	0.40	0.45	115
Opinionated	0.37	0.64	0.47	153
Positive	0.33	0.26	0.29	69
Negative	0.19	0.12	0.14	51
Neutral	0.21	0.11	0.14	84
None of the Above	0.82	0.70	0.76	20
Weighted Average	0.35	0.37	0.34	544

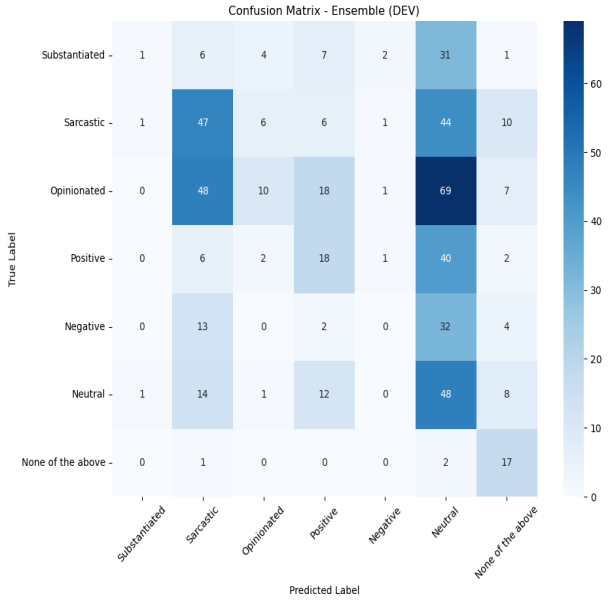

Figure 2: Confusion matrix of the proposed ensemble model. The matrix highlights frequent misclassification of minority classes such as Negative and Neutral into dominant classes like Opinionated.

to overlapping linguistic patterns and implicit sentiment expressions.

Figure 2 illustrates the confusion matrix of the proposed ensemble model. The confusion matrix reveals that minority classes such as Negative and Neutral are frequently misclassified as Opiniated, indicating residual bias towards dominant categories despite using of focal loss and oversampling techniques.

Although the proposed imbalance handling strategies improve robustness, overlapping linguistic patterns, implicit sentiment expressions, and sarcasm continue to present challenges for accurate classification (Lin et al., 2017). Frequent misclassifications between semantically similar categories further highlight the complexity of Tamil political sentiment analysis.

Overall, the results demonstrate that ensemble transformer architectures can provide moderate improvements in low-resource Tamil sentiment analysis tasks, emphasises the need for more advanced contextual and discourse-aware modeling approaches (Bergstra and Bengio, 2012). The implementation of the proposed model is publicly available at our GitHub repository: GitHub Code

Limitations

Despite its effectiveness, there are several limitations associated with the proposed system. The model performance is highly dependent on the size and quality of the dataset. Tamil social media data often contains informal language, dialectal variations, and code-mixed expressions that may not be fully represented in the training corpus.

In addition, the dataset exhibits severe class imbalance, where certain sentiment categories contain significantly fewer training examples. Although oversampling, balanced class weights, and focal loss help mitigate this issue, minority-class performance remains limited.

A sentence level sentiment classification is done on the proposed system and it does not explicit model discourse level context such as sarcasm, implicit meaning or conversational dependencies. I addition, transformer-based models such as XLM-RoBERTa and IndicBERT require substantial computational resources, which may limit deployment in resource-constrained environments (Conneau and Lample, 2019), (Kakwani et al., 2020).

Future work can explore context-aware architectures, larger domain-specific datasets, and advanced imbalance handling strategies to further improve performance in Tamil political sentiment analysis tasks.

References

B. K. Behera, A. Kumar, and S. Pradhan. 2021. Sentiment analysis in indian languages: A survey. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(3).

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, and et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Daniel Jurafsky and James H. Martin. 2023. *Speech and Language Processing*, 3 edition. Prentice Hall. Draft.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, and et al. 2020. Indicnlp suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual models for indian languages. In *Findings of EMNLP*, pages 4948–4961.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1145.

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)*, pages 2980–2988.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

M. Vegupatti, K. K. Ponnusamy, B. R. Chakravarthi, S. Rajiakodi, D. Thenmozhi, P. K. Kumaresan, and S. 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.