Trailblazer@DravidianLangTech 2026: A Comparative Study of TF-IDF SVM and XLM-RoBERTa for Political Multiclass Text Classification.

Shanthi Murugan , Anbuaruvi R , Anuradha C
Department of Artificial Intelligence and Data Science,
R.M.K. Engineering College
Kavaraipettai,Chennai, India
msi.ad@rmkec.ac.in, 230204.ad@rmkec.ac.in, 230042.ad@rmkec.ac.in

Abstract

The rapid growth of social media networks faces challenges in the classification of multilingual and code-mixed data. A task is shared by Political Multiclass Sentiment Analysis of Tamil X (Twitter) -DravidianLangTech@ACL 2026 to classify the political text.For the above task, we proposed solutions to compare a traditional machine learning and the transformer based model. First we developed a Baseline traditional Support vector Machine model using the TF-IDF features. To provide a stronger Indic-language baseline we consider the IndicBERT, a transformer model specifically designed for Indian Languages.IndicBERT improves contextual understanding of Tamil-English code-mixed political text . To capture the deeper information from the text we developed an XLM-RoBERTa model where we used minimal pre-processing technique. The Result shows us that the transformer-based performs well compared to the traditional baseline model with the macro F1 score of 0.3738. The Study highlights the importance of robust multi-class social media political text classification.

1 Introduction

The rapid expansion of social media platforms has transformed the way individuals communicate, share the opinions. Millions of posts are generated daily. The number of people who use social media is projected to grow exponentially, reaching nearly 1.2 billion users by 2029. In recent days, code-mixed text is used everywhere mainly in social media post. The task of sentiment analysis became difficult due to the usage of informal grammar, emojis, abbreviations, sarcasm, mixed language. The shared task on the Political Multiclass Sentiment Analysis of Tamil X (Twitter) - DravidianLangTech@ACL-2026 - aimed to address this challenge by identifying the types of political sentiments such as Substantiated, Sarcastic, Opinionated, Positive, Negative, Neutral, and None of the above. In Traditional machine learning approaches, faces struggle to capture the meaning of the text. Advancements in transformer-based models have significantly improved contextual language understanding.

In our participation, we focus on addressing the challenges of political sentiment classification through three primary contributions:

• We implemented three pre-processing strategies to bring the meaningful text understand by the model.

• We implement a Comparative framework between traditional ML model and the transformer based models

•We evaluated models using macro-F1 score to ensure balanced assessment across all sentiment categories rather than favoring majority classes.

Our code, developed for this shared task can be accessed here https://github.com/Anbuaruvi/Trailblazer.git

2 Related Works

The complexity of understanding and categorizing political sentiments has driven extensive research employing various languages, datasets, and methodologies. Research has been done to advance sentiment analysis in under-resourced code-mixed languages(Kumar et al., 2024) . Different machine learning models and Deep Learning models have been employed to classify sentiment of text(Barua et al., 2025). (Gandhar et al., 2024)conducted a comparative analysis of machine learning approaches such as Support Vector Machines (SVM) for sentiment analysis in Indian multilingual social media contexts using Twitter data. Their work highlights the challenges posed by mixed-language and code-mixed text data, and shows that SVM remains effective but limited for nuanced sentiment detection.Due to the inefficiency of machine learning models in extracting contex-

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 414–419

tual meanings, their works lack the ability to fully capture nuanced expressions and complex sentiment patterns.(Khurana and Prabhu, 2023)experiments on labeled Twitter data using transformer models such as BERT,ALBERT,RoBERTa and DeBERTa ,showing that fine-tuned transformer models significantly performs better than the classical machine learning in capturing sentiment in social media text.Another recent research in sentiment work includes multilingual sentiment analysis using the transformer model,this models support multiple Indian Languages(Kalla, 2025).Multilingual transformers have been explored for multiclass sentiment analysis in code-mixed data, effectively capturing contextual nuances in low-resource languages (Nazir et al., 2025). Collectively, these traditional and hybrid studies highlight the evolution of sentiment analysis research from feature-based and lexicon methods to more sophisticated context-aware models.

3 Task and Dataset Description

The Shared task focuses on the analysis of political multiclass sentiment of Tamil X (Twitter) post comments in Tamil language,which are annotated with seven categories .The main objective is to classify the sentiment from the given text.The Seven labels are as follows:

- **Substantiated** – Supported by facts, statistics, references, or specific details.
- **Sarcastic** – Sentiment based on sarcasm, mockery, or irony.
- **Opinionated** – Sentiment based on strong personal views or individual perspectives.
- **Positive** – Sentiment expressing support, approval, or appreciation.
- **Negative** – Sentiment based on criticism or dissatisfaction.
- **Neutral** – Sentiment based on general political information or factual updates without expressing opinions, emotions, or bias.
- **None of the above** – Sentiment that does not fit into any of the specified categories.

The distribution of the political sentiment classes across the training, development, test (Given), and Pred (Predicted) datasets is shown in Table 4. The training dataset exhibits a noticeable class imbalance, with the "Opinionated" category having the highest representation, significantly outnumbering other classes. In contrast, categories like "None" and "Substantiated" have relatively fewer samples, which is illustrated in Figure 3.

3.1 Tables and Visualization

Table 1: Label distribution across Train, Dev, Test (Given) and Predicted datasets.

Label	Train	Dev	Test	Pred
Substantiated	412	52	51	60
Sarcastic	790	115	106	102
Opinionated	1361	153	171	118
Positive	575	69	75	132
Negative	406	51	46	73
Neutral	637	84	70	36
None	171	20	25	23

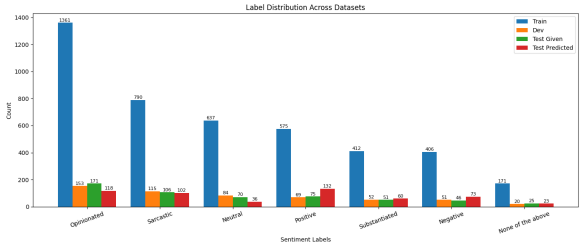

Figure 1: Distribution of sentiment labels across datasets.

4 Methodology

The Methods and Strategies used to classify the sentiment lables are detailed in this section.We implemented three different approaches:

1. A Baseline ML approach using TF-IDF with Support Vector Machine(SVM).(Ardiansyah et al., 2024)
2. An Indic-language transformer baseline using IndicBERT. (Kakwani et al., 2020)
3. A transformer-based approach using XLM-RoBERTa. (Liu et al., 2019)

4.1 Pre-Processing

4.1.1 Baseline SVM

The Traditional ML model requires careful preprocessing. The following are the applied preprocessing steps. Removal of URLs and Mentions like username, hastags , emojis , special symbols. This does not provide any useful meaning for sentiment analysis. Converting all the text to lowercase also removes the extra spaces.

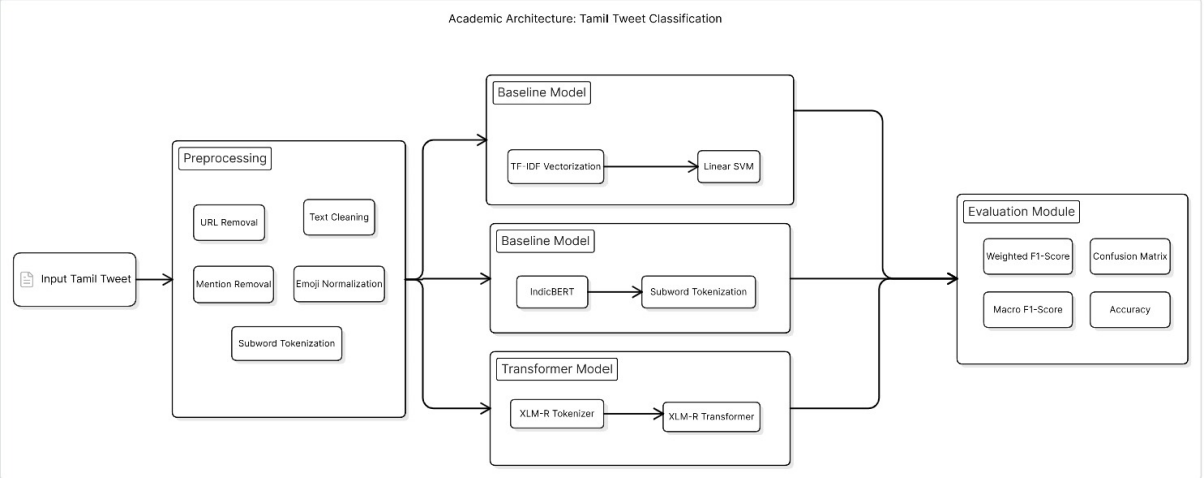

Figure 2: An Abstract view of Methodology

4.1.2 IndicBERT

The preprocessing steps applied for the IndicBERT include removal of URLs ,User mentions and unnecessary whitespace normalization. Unicode normalization was performed performed to maintain consistent Tamil text representation. Hashtags, punctuation, and casing information retained to preserve contextual and semantic meaning. The text was tokenized using subword tokenization for contextual embedding.

4.1.3 XLM-RoBERTa

Transformer-based models such as XLM-RoBERTa-base use subword tokenization and contextual embeddings.The steps performed are removal of URLs,user mentions, hastag symbol, whitespace normalization. No lowercasing or aggressive symbols removal is performed. The Minimal processing preserves semantics richness.

4.2 Modelling

4.2.1 Baseline SVM

The preprocessed text is converted into numerical feature vector using the Term Frequency-Inverse Document Frequency(TF-IDF). The n-gram range used here is unigram and bigrams using the TF-IDF features trained the Linear Support vector Machine.

The SVM Model is trained using the following to overcome the class imbalance:

1.Regularization parameter is set to C=1.0.

2.To handle class imbalance, the SVM was configured with class weight = "balanced", which assigns weights inversely proportional to class frequencies during training.

Table 2: IndicBERT Training Hyperparameters

Hyperparameter	Value
Learning Rate	2×10^{-5}
Batch Size	16
Epochs	5
Maximum Sequence Length	128
Warmup Ratio	0.1
Weight Decay	0.01

4.2.2 IndicBERT

The IndicBERT model was fine-tuned using the preprocessed Tamil X dataset for seven-class sentiment classification. The input text was tokenized using the IndicBERT tokenizer with padding and truncation applied to maximum sequence length of 128 tokens. Weighted cross-entropy loss and AdamW optimizer were used to address class imbalance and optimize training.

Table 3: XLM-RoBERTa Training Hyperparameters

Hyperparameter	Value
Learning Rate	2×10^{-5}
Batch Size	16
Epochs	6
Warmup Ratio	0.1
Weight Decay	0.01

4.3 XLM-RoBERTa

The transformer-based model used here is XLM-RoBERTa-base a mulitlingual transformer which is pre-trained on large languages .It consist of a token embedding layer, positional encodings, multiple

staked encoder blocks and a feed-forward networks and a final classification layer. The text input is processed using the XLM-RoBERTa tokenizer with a subword tokenization, truncation and padding to a maximum sequence of 128 tokens and attention to ensure uniform batch processing.

To Overcome class imbalance, class weights are computed from training data are incorporated into loss function and the model is optimized using the cross-entropy.

5 Results and Discussion

In this section, let us compare how the baseline SVM model, IndicBERT model and Transformer based XLM-RoBERTa perform on the given data. The effectiveness of the model is primarily assessed based on the F1 Score.The baseline model achieved the accuracy of 29.96 percentage with a Macro-F1 score of 0.2994 and the Weighted-F1 score of 0.2988.The low Macro-F1 score indicates that the model id difficulty in handling the imbalanced multi-class labels. Overall, the baseline model struggled to understand the semantic structure. IndicBERT perform better than traditional TF-IDF-based machine learning models because of its ability to capture contextual information and semantic relationships in Indic languages. The IndicBERT model achieved an accuracy of 0.3210 with a Macro-F1 score of 0.3345, outperforming the baseline TF-IDF + SVM model by effectively capturing contextual and semantic information present in Tamil-English code-mixed political text.

The transformer-based XLM-RoBERTa model performs better than the Baseline SVM.The model is trained using 6 epoches where result is The Accuracy 33.63 percentage,Macro-F1 score has 0.3640, Weighted-F1 score is 0.3395. The training log shows an improvement in each epoch from Epoch 1 to Epoch 6. Due to the stable increase overfitting did not occur. The transformer-based multilingual model XLM-RoBERTa learn contextual representation through self-attention mechanism,enabling better understanding of semantic relationships across multilingual text which makes transformer model work well compared to the baseline model. The Baseline SVM Model shows different performance across the class labels. In Class Label wise, the "None of the above", "Opinionated", "Sarcastic" shows higher Macro-F1 Score. Followed by The labels "Neutral", "Negative", "Positive" show the lower Macro-F1 score. The dataset contains significant class imbalance particularly with the "Opinionated" category dominating the training distribution while "Substantiated" and "None of the above" contains fever samples.

To Overcome this class imbalance, weighted cross-entropy loss is applied in the transformer model, also Macro-F1 is additionally added to evaluate the balanced performance across all the class labels.

From the baseline SVM Model to the Transformer-based model, the Macro-F1 score has increased approximately 6.5 percentage.

Table 4: Accuracy and Macro-F1 score for three different Models

Model	Accuracy	Macro-F1 score
SVM(Baseline)	0.2996	0.2994
IndicBERT	0.3210	0.3345
XLM-RoBERTa	0.3363	0.3640

Figure 3 represents the confusion matrix of the fine-tuned XLM-RoBERT model. The matrix provides a class label wise correct and incorrect predictions across all the seven labels. The highest number of correct prediction appears in the class label called **Opinionated** class which indicates that the transformer captures strong subjective expression. The class label **Sarcastic** shows robust performance with **56** correct predictions. The class label **None of the above** also classified the text with high precision. A significant number of Negative and Neutral labels are misclassified as Opinionated or Positive. The confusion matrix reveals that the transformer model successfully captures contextual relationships.

5.1 Error Analysis

Despite the improvement achieved by XLM-RoBERTa several challenges also remain.

- 1. Misclassification occurs between Negative, Neutral, Opinionated, and Positive sentiment classes. Class imbalance in the dataset causes the model to become biased toward majority classes such as Opinionated, leading to incorrect predictions for minority classes.
- 2. Code-mixed Tamil-English text creates difficulty in understanding contextual meaning and semantic relationships. Informal spelling variations and noisy social media text reduce the discriminative capability of the model.

- 3. Political social media discussions often contain emotions, sarcasm, ambiguous wording, and rhetorical expressions.

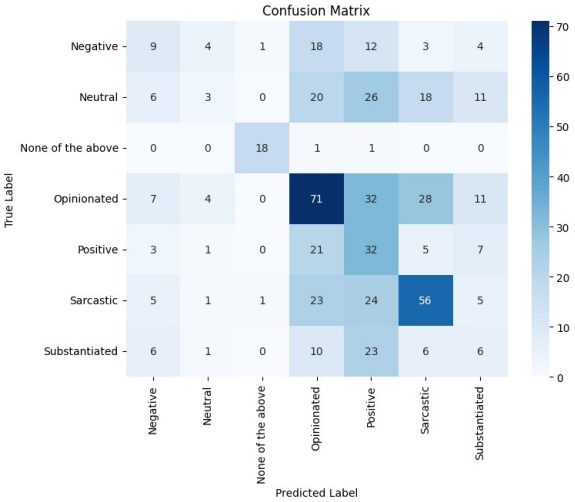

Figure 3: Confusion Matrix for XLM-RoBERTa

6 Conclusion

In this work, we proposed three different approaches for multiclass political sentiment classification: a traditional machine learning baseline model, an Indic-language transformer baseline, and a multilingual transformer-based model. We implemented TF-IDF with Linear SVM, IndicBERT, and fine-tuned XLM-RoBERTa for the classification of Tamil X (Twitter) political comments. The baseline SVM model extracts unigram and bigram TF-IDF features and combines them with a Linear Support Vector Machine for classification. However, the traditional model shows limitations in understanding contextual and semantic relationships, particularly for complex sentiment categories such as sarcasm and opinionated expressions.

In contrast, transformer-based models such as IndicBERT and XLM-RoBERTa generate contextual embeddings that improve semantic understanding of multilingual and code-mixed Tamil political text. IndicBERT achieved better performance than the baseline SVM model by effectively capturing Indic-language contextual information, while XLM-RoBERTa achieved the highest overall performance due to its stronger multilingual representation capability. To further address class imbalance, weighted cross-entropy loss was applied during transformer model training. The experimental results demonstrate that transformer-based multilingual approaches are more effective than traditional machine learning methods for low-resource political sentiment classification tasks.

References

R. Ardiansyah, H. Yuliansyah, and A. Yudhana. 2024. Twitter sentiment analysis of public space opinions using svm and tf-idf methods. *The Indonesian Journal of Computer Science*, 13(1).

Arupa Barua, Md Osama, and Ashim Dey. 2025. Cuet_novice@dravidianlangtech 2025: A bi-gru approach for multiclass political sentiment analysis of tamil twitter (x) comments. In *Proceedings of DravidianLangTech 2025*, India. DravidianLangTech.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Elizabeth Sherly, 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 DravidianLangTech@NAACL 2025*.

Abhishek Gandhar, Shashi Gandhar, S. B. Kumar, Arvind Rehalia, Prakhar Priyadarshi, and Mohit Tiwari. 2024. A comparative analysis of machine learning algorithms for sentiment analysis in indian social media.

Abirami Jayaraman, Aruna Devi Shanmugam, Dharunika Sasikumar, and Bharathi B. 2025. Analysisarchitects@dravidianlangtech 2025: Machine learning approach to political multiclass sentiment analysis of tamil. In *Proceedings of DravidianLangTech 2025*, Kalavakkam, Chennai, Tamil Nadu. Sri Sivasubramaniya Nadar College of Engineering.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, and 1 others. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Haribabu Kalla. 2025. Zero-shot multilingual sentiment analysis using transformer-based models. *IJERND*, 2(02).

Naman Khurana and Vibha Prabhu. 2023. Sentiment analysis using transformer models (bert, albert, roberta, and deberta) on a smile twitter dataset. *IEEE*.

Lavanya Sambath Kumar, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 2024.

Overview of second shared task on sentiment analysis in code-mixed tamil and tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 62–70, St. Julian's, Malta. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, and 1 others. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Muhammad Kashif Nazir, CM Nadeem Faisal, Muhammad Asif Habib, and Haseeb Ahmad. 2025. Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages. *IEEE Access*.

Billodal Roy, Souvik Bhattacharyya, Pranav Gupta, and Niranjan Kumar M. 2025. Lexilogic@dravidianlangtech 2025: Political multiclass sentiment analysis of tamil x (twitter) comments and sentiment analysis in tamil and tulu. In *Proceedings of DravidianLangTech 2025*. Lowe's.

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.