Hermes@DravidianLangTech 2025: Sentiment Analysis of Dravidian Languages using XLM-RoBERTa

Emmanuel George P[1], **Ashiq Firoz**[1], **Madhav Murali**[1]
Siranjeevi Rajamanickam[2] , **Balasubramanian Palani**[3]
[1]Department of Computer Science and Engineering, IIIT Kottayam
[2]Lecturer, Dept of Computer Engineering, Govt. Polytechnic College-Trichy
[3]Assistant Professor, Indian Institute of Information Technology Kottayam
{emmanuel22bcs104,ashiq22bcd13,madhav22bcs50,pbala}@iiitkottayam.ac.in,rajasiranjeevi@gmail.com

Abstract

Sentiment analysis, the task of identifying subjective opinions or emotional responses, has become increasingly significant with the rise of social media. However, analyzing sentiment in Dravidian languages such as Tamil-English and Tulu-English, presents unique challenges due to linguistic code-switching (where people tend to mix multiple languages) and non-native scripts. Traditional monolingual sentiment analysis models struggle to address these complexities effectively. This research explores a fine-tuned transformer model based on the XLM-RoBERTa model for sentiment detection. It utilizes the tokenizer from the XLM-RoBERTa model for text preprocessing. This research is based on our work for the Sentiment Analysis in Tamil and Tulu DravidianLangTech@NAACL 2025 competition. We received an F1-score of 71% for the Tulu dataset and 60% for the Tamil dataset, which placed us third in the competition.

1 Introduction

Sentiment analysis plays a pivotal role in understanding subjective opinions and emotional responses in text. With the growing prominence of social media, the demand for sentiment analysis on user-generated content has surged. However, social media texts often include code-mixed content, where multiple languages are blended within a single sentence. This phenomenon is particularly prevalent in multilingual communities, such as those speaking Dravidian languages, where Tamil-English and Tulu-English code-mixing is common. These texts often incorporate linguistic code-switching and non-native scripts, adding layers of complexity to the task of sentiment analysis.

The novelty of this research lies in its focus on sentiment analysis for low-resource code-mixed Dravidian languages, an area that has remained underexplored despite the growing demand for multilingual natural language processing (NLP) solutions. Additionally, this study is based on a fine-tuned XLM-RoBERTa (Liu et al., 2019) model, providing new insights into their strengths and limitations in handling linguistic complexities like code-switching and non-native scripts. The fine-tuned transformer approach, in particular, demonstrates a superior ability to address these challenges by effectively leveraging its contextual understanding capabilities.

The dataset (Chakravarthi et al., 2020b)used in this study contained four and five sentiment classes in Tamil and Tulu language, respectively. The XLM-RoBERTa model (Conneau et al., 2020)was trained using these datasets for the multi-class classification task. Later the macro F1-score metric of the validation dataset was used to evaluate the model performance. This approach secured a third-place ranking in the competition(Durairaj et al., 2025).

A comparison of the results of XLM-RoBERTa model with tradition machine learning models like Logistic Regression and Random Forest in combination with the TF-IDF vectorizer for tokenization was done and this comparison was extended to the transformer models BERT base (Devlin et al., 2019) and RoBERTa base (Palani and Elango, 2023). The XLM-RoBERTa model provided a better result in comparison with all these models also.

2 Literature Survey

Ahmad et al. (2022) (Ahmad et al., 2022) provide a comprehensive review of machine learning techniques for sentiment analysis in code-mixed and switched text, emphasizing the challenges posed by bilingual and multilingual expressions in Indian social media contexts. Chakravarthi et al. (2020) (Chakravarthi et al., 2020a) introduced a gold standard corpus for Malayalam-English code-mixed text, which serves as a benchmark for sentiment

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 408–412

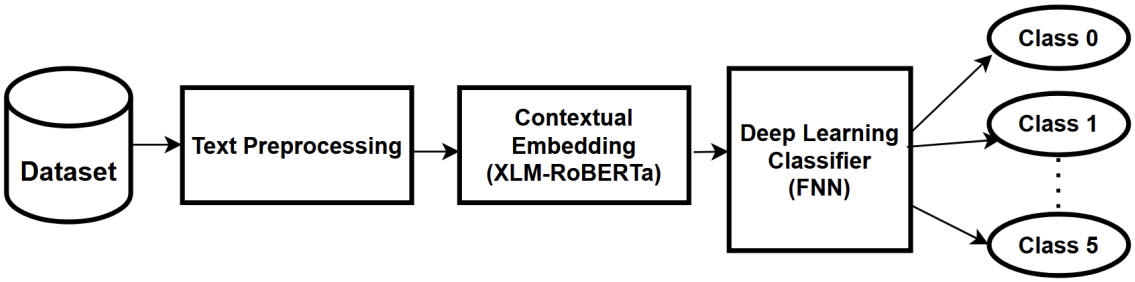

Figure 1: Architecture of the proposed model for sentiment prediction

analysis tasks with high inter-annotator agreement. The Language Technologies Research Center at IIIT Hyderabad studied (Mishra et al., 2018) sentiment detection in Hindi and Bengali using a voting classifier with an SVM model and TF-IDF vectorization, achieving accuracies of 56% for Hindi and 52% for Bengali. RANLP 2023 explored sentiment analysis for code-mixed Tamil and Tulu texts, (Hegde et al., 2023) reporting macro-average F1 scores of 0.32 for Tamil and 0.542 for Tulu, underscoring the growing interest in Dravidian sentiment analysis and the need for further advancements.

3 Methodology

We followed a series of steps to develop the model for sentiment prediction, which included dataset preprocessing, tokenization, training, validation, prediction, and model evaluation. Each of these steps is discussed in detail in this section, with visual representation provided in Figure 1.

3.1 Problem Definition

The sentiment analysis task involves two distinct datasets: one for Tamil and another for Tulu. Given a Tamil dataset $T = \{t_1, t_2, \ldots, t_m\}$ consisting of m social media comments, each comment $t_i \in T$ is associated with one of the following class labels: Positive, Negative, Mixed Feelings, or Unknown State. Separately, the Tulu dataset $U = \{u_1, u_2, \ldots, u_k\}$ consists of k social media comments, each labeled with one of five categories: Not Tulu, Positive, Negative, Neutral, or Mixed. Classification models $f_T : T \rightarrow y$ and $f_U : U \rightarrow y$ are defined and trained to predict the corresponding class label for each comment in their respective datasets.

3.2 Data Preprocessing

Data processing prepares the dataset for training the machine learning model to detect the different sentiments.The dataset had multiple entries with Nan values in it. We had to remove these values from the dataset and we went on to map the class labels to the integer values using a label map.

3.3 Tokenization

Converted the textual data to numbrs using the process of tokenization and the tokenizer we used was the XLM-RoBERTa tokenizer which would align perfectly with our classifier model. Similarly the RoBERTa base and BERT base models (Palani and Elango, 2023) utilized their respective tokenizers and the machine learning models utilized the TF-IDF vectorizer (Kanta and Sidorov, 2023) with max_features set to 5000.

3.4 Model Architecture

The model we used here is the XLM-RoBERTa (Conneau et al., 2020) model and feed forward networks(FFN). We set the problem_type parameter to 'single_label_classification,' meaning that each data point will be assigned to only one of the target classes and num_labels to the number of classes in the dataset. (Liu et al., 2019) Both RoBERTa base and BERT base(Devlin et al., 2019) where also set to the same parameters. Incase of Logistic Regression it had max_iter set to 1000 and for Random Forest n_estimators was set to 200 .

3.5 Model Training

The model was trained using the AdamW optimizer (learning rate: 2 x 10^{-5}, weight decay: 0.01) with a linear warm-up schedule. A batch size of 16 was used, and training ran for 10 epochs with early stopping (patience: 3). Cross-Entropy Loss was

Table 1: Samples of different class labels in the dataset

Tamil		Tulu	
Positive	2020 முதல் வெற்றி மாஸ் வெறித்தனமான	Positive	ಅಣ್ಣ ಮಸ್ತ್ ಖುಷಿ ಆಪುಂಡು ಇರೆನ ಶೋ ತೂವರೆ
Negative	திமிர் பிடித்த திருநங்கைகள்	Negative	ಎಂಚಿ ಸಾವುದ ಪುಕುಳಿಯ
Mixed_feelings	சில்லறை தராமல் எடுத்துக் கொண்டு போனார்	Mixed	ಅಂಬಾನಿ ತುಂಡಾ ಸೈತೆ ಪೊವೇ
unknown_state	இந்த படத்தை டிவில பத்து டே 2022 ல	Neutral	ನೀರ್ ದ ಮಹತ್ವನ್ ತೇರಿಲೆ
-	-	Not Tulu	Congratulations Mohan sir

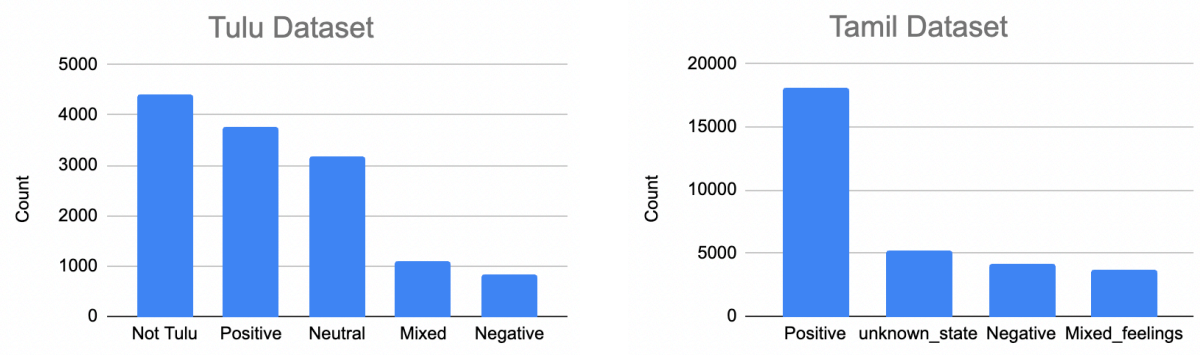

Figure 2: Frequency distribution of classes in the training dataset

applied, and overfitting was monitored using the weighted average F1-score on the test set. Transformer models followed a similar training setup, while machine learning models used the fit function.

3.6 Model Evaluation

The model was evaluated using the validation dataset. For each datapoint, the model predicted that it belonged to a predefined class label. Now, the different evaluation metrics were found and compared on the basis of the models.

4 Experiment

This section gives us a comprehensive study of the experimental setup and the different performance metrics utilized in this study.

4.1 Experimental Setup

The study uses machine learning and deep learning approaches to classify sentiments. These models were implemented and tested in Python programming language using PyTorch, Transformers, Pandas, Scikit-learn, and Tqdm for data processing, model training, and analysis. Hugging Face tokenizers handled text preprocessing and training was performed on CUDA-enabled GPUs, leveraging PyTorch for efficient deep learning and Hugging Face API for streamlined access to the models.

4.2 Dataset Description

This study utilizes multilingual sentiment analysis datasets in Tamil and Tulu, including cleaned subsets of Tamil data. The dataset(Hegde et al., 2022) contains training and validation sets and the frequency distribution of different class labels in the training dataset are shown in Figure 2. Some samples of each of these class labels are also shown in Table 1 and a summary of the datasets is shown in Table 2.

The dataset incorporates both code-mixed and romanized data points and samples of code-mixed and romanized data points are as given below:

- Code-mixed: படம் வெற்றிபெற நாடார் சமுகத்தினர் சார்பாக வாழ்த்துகள்
- Romanized: kandipa nama Ella records um break panuvom

Table 3: Performance comparison of XLM-RoBERTa model with other standard models on Tamil and Tulu datasets

Model	Tulu Dataset				Tamil Dataset			
	Acc.	Prec.	Rec.	Macro F1	Acc.	Prec.	Rec.	Macro F1
XLM-RoBERTa	0.71	0.63	0.59	0.59	0.65	0.51	0.49	0.49
RoBERTa base	0.67	0.58	0.55	0.54	0.63	0.48	0.43	0.44
BERT base	0.69	0.59	0.59	0.59	0.64	0.49	0.48	0.48
TF-IDF (LR)	0.64	0.55	0.57	0.55	0.53	0.44	0.48	0.45
TF-IDF (SVM)	0.64	0.54	0.51	0.51	0.63	0.49	0.39	0.41

Table 2: Summary of datasets

Dataset	Training	Testing
Tamil	31,122	3,843
Tulu	13,308	1,643

4.3 Evaluation Metrics

The performance of the model is evaluated using Accuracy, Macro Precision, Macro Recall, and Macro F1-score.

$$\text{Accuracy (Acc.)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Macro Precision (Prec)} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i} \quad (2)$$

$$\text{Macro Recall (Rec.)} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \quad (3)$$

$$\text{Macro F1} = \frac{2 \times \text{Macro Precision} \times \text{Macro Recall}}{\text{Macro Precision} + \text{Macro Recall}} \quad (4)$$

Here, TP, TN, FP, and FN represent the true positives, true negatives, false positives, and false negatives, respectively, and N is the number of classes.

5 Results

The best results for Tamil and Tulu sentiment analysis were given by the XLM-RoBERTa model, with a macro F1-score of 59% for the Tulu dataset and 49% for the Tamil dataset. XLM-RoBERTa model was actually trained on multi-lingual text and this is the reason why it has an edge over the other models. The details of the macro of the evaluation metrics and accuracy are as shown in the Table 3. The best macro average of f1-score we received was 59% and 49% for testing dataset of Tulu and Tamil language respectively.

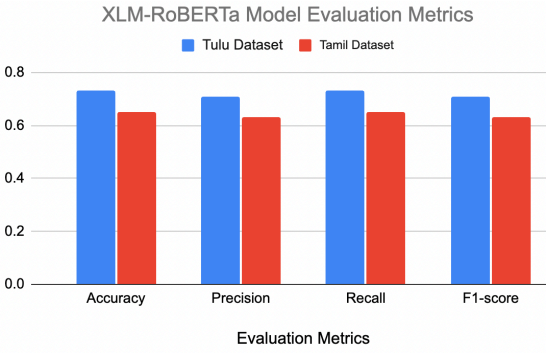

Figure 3: Evaluation metrics of XLM-RoBeRTa model.

From Figure 3 we can infer that the Tulu dataset is performing better than the Tamil dataset, this is because of the irregular frequency of the class labels in the Tamil dataset. As you can see in the Figure 2 the number of positive classes in the dataset is very high compared to the rest of the classes causing class imbalance thus leading to lesser accuracy.

6 Conclusion and Future Directions

Sentiment analysis in code-mixed text faces challenges like class imbalance and code-switching complexities, with fine-tuned transformers helping but limited by multilingual resource gaps. The Tamil dataset shows significant class imbalance, affecting model performance, with XLM-RoBERTa achieving a macro F1-score of 59% for Tulu and 49% for Tamil. Future work could explore re-sampling, cost-sensitive learning, and advanced data augmentation to address imbalance, while developing comprehensive lexical resources would enhance transformer-based sentiment analysis for multilingual communities.

References

Gazi Ahmad, Jimmy Singla, Anis Ali, Aijaz Reshi, and Anas A. Salameh. 2022. Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus - a comprehensive review. *International Journal of Advanced Computer Science and Applications*, 13.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. 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, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. 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, Marseille, France. 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. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Asha Hedge, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Kalyanasundaram Krishnakumari, Charumathi Rajkumar, Poorvi Shetty, and Harshitha S Kumar. 2025. Overview of the Shared Task on Sentiment Analysis in Tamil and Tulu: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, Lavanya S K, Thenmozhi D., Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in Tamil and Tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Selam Kanta and Grigori Sidorov. 2023. Selam@DravidianLangTech:sentiment analysis of code-mixed Dravidian texts using SVM classification. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 176–179, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Pruthwik Mishra, Prathyusha Danda, and Pranav Dhakras. 2018. Code-mixed sentiment analysis using machine learning and neural network approaches. *Preprint*, arXiv:1808.03299.

Balasubramanian Palani and Sivasankar Elango. 2023. Ctrl-fnd: content-based transfer learning approach for fake news detection on social media. *International Journal of System Assurance Engineering and Management*, 14(3):903–918.