SSNTrio@DravidianLangTech 2025: Sentiment Analysis in Dravidian Languages using Multilingual BERT

Bhuvana J
Sri Sivasubramaniya Nadar College of Engineering
bhuvanaj@ssn.edu.in

Mirnalinee T T
Sri Sivasubramaniya Nadar College of Engineering
MirnalineeTT@ssn.edu.in

Diya Seshan
Sri Sivasubramaniya Nadar College of Engineering
diya2210208@ssn.edu.in

Rohan R
Sri Sivasubramaniya Nadar College of Engineering
rohan2210124@ssn.edu.in

Avaneesh Koushik
Sri Sivasubramaniya Nadar College of Engineering
avaneesh2210179@ssn.edu.in

Abstract

This paper presents an approach to sentiment analysis for code-mixed Tamil-English and Tulu-English datasets as part of the DravidianLangTech@NAACL 2025 shared task. Sentiment analysis, the process of determining the emotional tone or subjective opinion in text, has become a critical tool in analyzing public sentiment on social media platforms. The approach discussed here uses multilingual BERT (mBERT) fine-tuned on the provided datasets to classify sentiment polarity into various predefined categories: for Tulu, the categories were positive, negative, not_tulu, mixed, and neutral; for Tamil, the categories were positive, negative, unknown, mixed_feelings, and neutral. The mBERT model demonstrates its effectiveness in handling sentiment analysis for code-mixed and resource-constrained languages by achieving an F1-score of 0.44 for Tamil, securing the 6th position in the ranklist; and 0.56 for Tulu, ranking 5th in the respective task.

1 Introduction

Sentiment analysis involves identifying subjective opinions or emotional responses related to a specific topic. Over the past two decades, it has gained significant attention in both academia and industry. The demand for sentiment detection in social media texts, especially those containing code-mixing in Dravidian languages, has been steadily increasing.

The DravidianLangTech@NAACL 2025 shared task (Durairaj et al., 2025) focuses on analyzing sentiment in code-mixed Tamil-English and Tulu-English text collected from social media platforms like YouTube (S. K. et al., 2024). Both languages represent diverse linguistic characteristics, with Tulu being particularly underrepresented in computational linguistic research due to its limited annotated datasets. The datasets provided reflect real-world scenarios, including short, informal, and noisy social media posts, accompanied by significant class imbalance across sentiment categories (Hegde et al., 2023).

This study presents a sentiment analysis system leveraging Multilingual BERT (mBERT). The model was fine-tuned on the provided datasets to classify sentiment into various predefined categories. For Tamil, these categories included positive, negative, unknown, mixed_feelings, and neutral, while for Tulu, they comprised positive, negative, not_tulu, mixed, and neutral. Additionally, significant class imbalance was encountered, which was addressed using upsampling techniques.

The system achieved an F1-score of 0.44 for Tamil, securing 6th place, and an F1-score of 0.56 for Tulu, earning 5th place in the competition. These results demonstrate the effectiveness of the approach in tackling sentiment analysis for code-mixed and low-resource Dravidian languages, while also contributing to the development of robust methods for analyzing code-mixed social media text.

2 Related Works

Sentiment analysis plays a vital role in gauging public opinion on social media. Over the years, various approaches have been successful in this domain, ranging from traditional machine learning techniques to more recent deep learning methods. Initially, models like Support Vector Machines (SVMs) and Naive Bayes were widely used for sentiment classification, leveraging handcrafted features such as term frequency and n-grams (Sugitomo et al., 2021). However, the advent of deep learning revolutionized sentiment analysis, with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks improving the ability to capture sequential dependencies in text (Srinivas et al., 2021).

More recently, transformer-based models like BERT and its multilingual variants, such as mBERT, have set new benchmarks in sentiment analysis (Krasitskii et al., 2025). Additionally, stud-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 418–422

ies on Dravidian languages have highlighted the challenges of sentiment analysis in code-mixed languages, with recent works exploring the use of multilingual models and transfer learning to address these issues (Perera and Caldera, 2024).

Despite advances, challenges remain, especially for underrepresented languages like Tulu, which lack annotated resources. Solutions like data augmentation, upsampling, and fine-tuning transformer models are needed to improve performance on code-mixed, low-resource datasets. This study advances sentiment analysis for Dravidian languages, focusing on Tamil-English and Tulu-English social media text.

3 Methodology

The implementation details and source code are available on github.[1]

3.1 Dataset Description

The datasets provided for this task consisted of train, dev, and test splits. For both Tamil and Tulu, the train and dev datasets had text and category as labels, while the test dataset only contained ID and text as labels.

The Tamil train dataset consisted of 31,122 rows, and the dev dataset consisted of 1,643 rows, distributed as shown in Table 1 and Figure 1 (Chakravarthi et al., 2020).

The Tulu train dataset consisted of 13,301 rows, and the dev dataset consisted of 1,643 rows, as shown in Table 2 and Figure 2 (Hegde et al., 2022).

The test dataset consisted of 3,459 rows for Tamil and 1,479 rows for Tulu.

The datasets were sourced from social media platforms like YouTube, containing real-world, informal, and noisy text. These texts were code-mixed, presenting challenges such as transliteration, spelling variations, and grammatical inconsistencies. Furthermore, significant class imbalance was observed in both Tamil and Tulu datasets, which required careful preprocessing and handling during model training.

3.2 Data Preprocessing

In text classification tasks, effective data preprocessing is crucial to ensure that the model can learn meaningful patterns from the raw data. This involves cleaning, transforming, and structuring raw

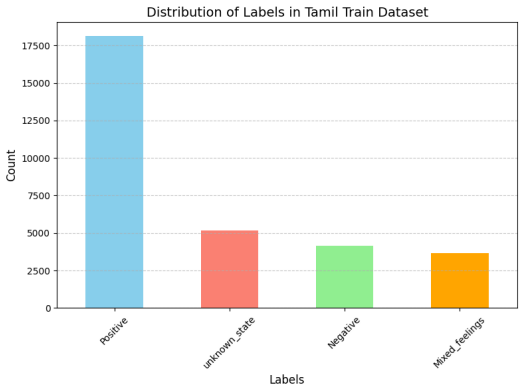

Figure 1: Distribution of Training Labels for Tamil

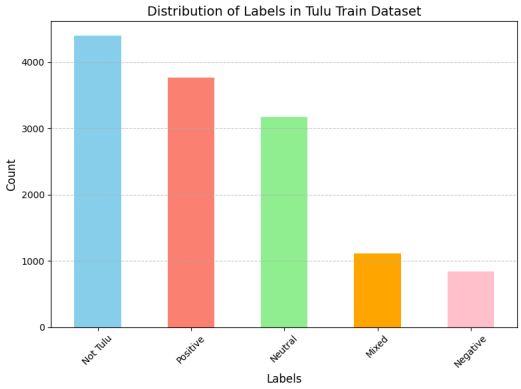

Figure 2: Distribution of Training Labels for Tulu

data into a format suitable for analysis or model training.

One of the key preprocessing steps involved eliminating punctuation marks and special characters. This removal helps reduce noise and inconsistencies in the data, allowing the model to focus on the core content of the text. By minimizing irrelevant elements, the model can better identify the essential features for sentiment analysis.

Tokenization is another crucial preprocessing step in this approach. The input text is tokenized using the BertTokenizer from the pre-trained mBERT model (bert-base-multilingual-cased). The tokenizer splits the text into smaller units, or tokens, that the model can process effectively. These tokens are then padded to a maximum length of 128 and truncated when necessary to fit within the model's input size. This ensures uniformity in input lengths across different samples, allowing for efficient training and model inference.

3.3 Synthetic Sample Generation

A significant challenge encountered during exploratory data analysis was class imbalance, where

[1] https://github.com/DiyaSeshan/DravidianLangTech2025-Sentiment-Analysis/tree/main

	Positive	Negative	Unknown State	Mixed Feelings
Training Data	18,145	4,151	5,164	3,662
Validation Data	2,272	480	619	472

Table 1: Distribution of Sentiment Labels in Tamil Dataset

	Positive	Negative	Neutral	Mixed	Not Tulu
Training Data	3,769	843	3,175	1,114	4,400
Validation Data	470	118	368	143	543

Table 2: Distribution of Sentiment Labels in Tulu Dataset

certain sentiment categories were underrepresented in the dataset. For example, in Tamil, the labels other than Positive were all significantly underrepresented, with Unknown State, Negative, and Mixed Feelings being much less frequent. Similarly, in Tulu, the Mixed and Negative sentiment labels had considerably lower representation compared to the Positive, Neutral, and Not Tulu categories.

To mitigate this issue, RandomOverSampler was used from the unbalanced-learn library. This technique randomly duplicates samples from minority classes, creating a more balanced class distribution. This balances the class distribution, helping to prevent the model from becoming biased towards the majority class, which could negatively impact performance (Permataning Tyas et al., 2023). In this study, RandomOverSampler saturated the number of rows for all labels to match the maximum class size, resulting in 18,145 rows per label in the Tamil training dataset and 4,400 rows per label in the Tulu training dataset.

3.4 Proposed Model

After performing data preprocessing, tokenization, and random oversampling, the dataset was split into training and validation subsets to allow effective evaluation of model performance. Three transformer-based models were experimented with: mBERT, Tamil BERT (from Hugging Face), and IndicBERT (from AI4Bharat). Among the models tested, mBERT demonstrated superior performance due to its multilingual capabilities, making it suitable for both Tamil and Tulu datasets. mBERT is a transformer-based language model pre-trained on 104 languages, making it highly effective for handling code-mixed text and providing robust performance across a wide range of languages, including Tamil and Tulu.

During training, the following hyperparameters were used:

- **Batch size**: Determines the number of training samples used in one forward/backward pass. Set to 16.
- **Maximum token length**: Defines the maximum number of tokens in each input sequence, ensuring the model processes inputs efficiently. Set to 128.
- **Optimizer**: AdamW, which combines the benefits of Adam with weight decay for regularization, often improving generalization.
- **Learning rate**: Controls how much to change the model's weights with respect to the loss gradient during training. Set to 3e-5.

Class weights were computed using compute_class_weight to adjust the loss function. The model was trained for 10 epochs, and evaluation metrics such as accuracy and classification reports were generated to assess performance on the validation set.

While Tamil BERT and IndicBERT also performed reasonably well, they did not match the performance of mBERT on this specific task. Tamil BERT is tailored for the Tamil language but lacks the multilingual capabilities of mBERT. IndicBERT, while promising for Indian languages, likely faced challenges with the code-mixed nature of the dataset. Given these comparisons, mBERT's multilingual training and robustness in handling diverse linguistic inputs made it the optimal choice for sentiment analysis in this study.

4 Results

In the training phase, the model achieved F1-scores of 0.8055 for Tamil and 0.8173 for Tulu, indicating that it successfully captured the sentiment patterns within the training data despite the challenges

posed by the code-mixed nature of the text and the class imbalances present in both languages. Moreover, the model demonstrated strong performance in identifying mixed feelings, unknown states, and neutral sentiments, effectively distinguishing these complex sentiment categories amidst the diverse and noisy social media data.

In the testing phase as well, the model demonstrated notable performance for both Tamil and Tulu. For Tamil, the F1-score achieved was 0.4461, ranking 6th in the ranklist. For Tulu, the model performed slightly better, achieving a F1-score of 0.569, ranking 5th in the ranklist. Despite the challenges posed by Tulu, a low-resource and underrepresented language with limited annotated data and scarce contextual information, the model's performance was commendable. The successful handling of Tulu sentiment analysis showcases the effectiveness of the discussed approach in overcoming barriers such as data sparsity and linguistic underrepresentation.

These results indicate a promising direction for future improvements in sentiment analysis for Dravidian languages.

Metric	Precision	Recall	F1-score
Positive	0.86	0.65	0.74
Negative	0.90	0.94	0.92
unknown_state	0.81	0.94	0.87
Mixed_feelings	0.89	0.92	0.91

Table 3: Performance Metrics for Sentiment Analysis of Tamil

Metric	Precision	Recall	F1-score
Positive	0.77	0.82	0.79
Negative	0.89	0.99	0.93
Neutral	0.87	0.69	0.77
Mixed	0.81	0.93	0.87
Not Tulu	0.90	0.79	0.84

Table 4: Performance Metrics for Sentiment Analysis of Tulu

5 Conclusion

In conclusion, this study presents a successful approach to sentiment analysis for code-mixed Tamil-English and Tulu-English datasets, demonstrating the effectiveness of mBERT in handling the complexities of code-switching and low-resource languages. Despite the challenges posed by noisy social media data and class imbalance, the model achieved promising results, achieving competitive F1-scores across sentiment classes.

The results highlight the model's strength in capturing sentiment nuances, especially for mixed feelings, unknown state, and neutral categories. They reinforce the potential of transformer-based models in advancing sentiment analysis for underrepresented languages and lay the foundation for future improvements through enhanced data preprocessing, augmentation, and fine-tuning strategies.

6 Future Enhancements

While transformer-based models like mBERT have shown promising results in sentiment analysis for code-mixed languages, several areas remain open for improvement. Future research could explore domain-adaptive pretraining, where models are fine-tuned on social media-specific corpora to enhance their understanding of informal and code-mixed text.

Additionally, incorporating external linguistic resources, like code-mixed lexicons and transliteration models, could help improve accuracy, especially for low-resource languages like Tulu. These resources can bridge the gap in language understanding and provide better context for sentiment analysis in code-mixed data.

Finally, integrating multimodal sentiment analysis by combining textual, visual, and audio cues from social media posts could provide a more comprehensive understanding of sentiment, enhancing real-world applications in social media monitoring and customer sentiment analysis.

7 Limitations

The discussed approach relies on a predefined dataset structure, which may limit its generalizability to a variety of real-world scenarios where data distributions differ significantly. Furthermore, the model's computational complexity increases with input length, which reduces its scalability for very long texts without substantial optimization.

Another limitation is the dependence on high-performance hardware, such as GPUs, for efficient training and inference, particularly since the datasets used in this study were extremely large. This could pose challenges for deployment in resource-constrained environments, where access to such computational resources is limited.

References

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. 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.

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*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Mikhail Krasitskii, Olga Kolesnikova, Liliana Chanona Hernandez, Grigori Sidorov, and Alexander Gelbukh. 2025. Comparative approaches to sentiment analysis using datasets in major european and arabic languages. In *Artificial Intelligence and Big Data Trends 2025*, AIBD, page 137–150. Academy Industry Research Collaboration Center.

Anne Perera and Amitha Caldera. 2024. Sentiment analysis of code-mixed text: A comprehensive review. *JUCS - Journal of Universal Computer Science*, 30(2):242–261.

Salsabila Mazya Permataning Tyas, Riyanarto Sarno, Agus Tri Haryono, and Kelly Rossa Sungkono. 2023. A robustly optimized bert using random oversampling for analyzing imbalanced stock news sentiment data. In *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pages 897–902.

Lavanya S. K., Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, and Rajkumar Charmathi Kumaresan, Prasanna Kumar. 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*, Malta. European Chapter of the Association for Computational Linguistics.

Akana Srinivas, Ch Satyanarayana, Ch Divakar, and Katikireddy Sirisha. 2021. Sentiment analysis using neural network and lstm. *IOP Conference Series: Materials Science and Engineering*, 1074:012007.

Jason Cornelius Sugitomo, Nathaniel Kevin, Nayra Jannatri, and Derwin Suhartono. 2021. Sentiment analysis using svm and naïve bayes classifiers on restaurant review dataset. In *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, volume 1, pages 100–108.