YenLP_CS@DravidianLangTech 2025: Sentiment Analysis on Code-Mixed Tamil-Tulu Data Using Machine Learning and Deep Learning Models

Raksha Adyanthaya
Department of Computer science,
Yenepoya Institute Of Arts,
Science, Commerce and Management,
Yenepoya (Deemed to be University),
Balmatta, Mangalore
rakshaadyanthaya11@gmail.com

Rathnakar Shetty P
Department of Computer science,
Yenepoya Institute Of Arts,
Science, Commerce and Management,
Yenepoya (Deemed to be University),
Balmatta, Mangalore
rathnakar.sp@gmail.com

Abstract

The sentiment analysis in code-mixed Dravidian languages such as Tamil-English and Tulu-English is the focus of this study because these languages present difficulties for conventional techniques. In this work, We used ensembles, multilingual Bidirectional Encoder Representation (mBERT), Bidirectional Long Short Term Memory (BiLSTM), Random Forest (RF), Support Vector Machine (SVM), and preprocessing in conjunction with Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec feature extraction. mBERT obtained accuracy of 64% for Tamil and 68% for Tulu on development datasets. In test sets, the ensemble model gave Tamil a macro F1-score of 0.4117, while mBERT gave Tulu a macro F1-score of 0.5511. With regularization and data augmentation, these results demonstrate the approach's potential for further advancements.

Keywords: Code-mixed, Classification, Dravidian language, Low resource language, Sentiment analysis, mBERT, Tamil, Tulu

1 Introduction

The internet's explosive growth has resulted in a proliferation of user-generated content on forums, blogs, social media, and e-commerce sites (Nazir et al., 2025). According to (Hande et al., 2020) there are more than 2.5 million speakers of Tulu in parts of Karnataka and Kerala, while Tamil, one of the oldest classical languages, is the official language of Tamil Nadu and Pondicherry. Online reviews and social media content are examples of textual data that can be accurately analyzed using sentiment analysis, a crucial Natural Language Processing (NLP) task (Fauzi, 2018).

Code-mixing is the process of combining several languages at different levels, such as words, phrases, or sub-words, within a single text. A number of factors contribute to code-mixing, such as social status, the speaker and their conversation partner, language, social community, bilingualism, and the circumstance or setting. These factors significantly influence code-mixing. Code-mixing often occurs when a term or phrase is unavailable in a given language, forcing people to use words or phrases from their own tongue to improve the receiver's comprehension(Ehsan et al., 2023). Sentiment analysis, which has applications in business, government and finance, is the automatic identification and interpretation of emotions or opinions in text (Wang, 2023). Since raw text cannot be directly processed by machine learning classifiers, feature extraction is crucial to convert text into numerical representations. The gap between unstructured text and machine learning algorithms is filled by methods such as Word2Vec, which records semantic context, and TF-IDF, which evaluates word relevance (Al-Kharboush and Al-Hagery, 2021). The text was classified according to sentiment using machine learning and deep learning models, and the results were analyzed to determine how well each strategy performed.

2 Literature Review

In Tamil, sentiment analysis has been thoroughly studied using conventional and contemporary deep learning techniques (Ehsan et al., 2023), while Tulu, a Dravidian language with few resources, has received little attention. Cultural quirks, idioms, sarcasm, and distinct syntax make it challenging to reveal hidden sentiments in under-resourced and code-mixed languages like Tamil, Tulu, and Kannada, even with high-quality datasets (Hussein, 2018). Furthermore, model generalization is impeded by class imbalances and limited datasets (Hande et al., 2020).

To address these issues, research has experimented with methods such as kNN, RF, and SVM, along with GridSearch for fine-tuning hyperparameters. The DravidianLangTech's Second Shared

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 366–370
May 3, 2025 ©2025 Association for Computational Linguistics

Task (EACL-2024) on analyzing sentiment in code-mixed Tamil and Tulu, as detailed in the work of Kumar et al. (2024), produced macro F1-scores of 0.260 for Tamil and 0.584 for Tulu. This underscores the importance of ensemble learning and optimization. The pre-training and fine-tuning techniques of BERT, as presented by Kenton and Toutanova (2019), have revealed promising outcomes.

Among the models created for the DravidianLangTech Shared Task (EACL-2024) (Prathvi et al., 2024) are an ensemble model that combines RF, kNN, SGD, and Logistic Regression with hard voting, as well as a LinearSVC model. Both models employ GridSearch for tuning and TF-IDF and CountVectorizer for n-gram extraction. With macro F1-scores of 0.260 for Tamil and 0.550 for Tulu, the ensemble model outperformed LinearSVC and showed promise for code-mixed sentiment analysis.

3 Methodology

Dataset preprocessing, feature extraction, machine learning and deep learning model creation, ensemble classification, and assessment are the main steps of the approach employed in this work. The proposed methodology is presented as Figure 1

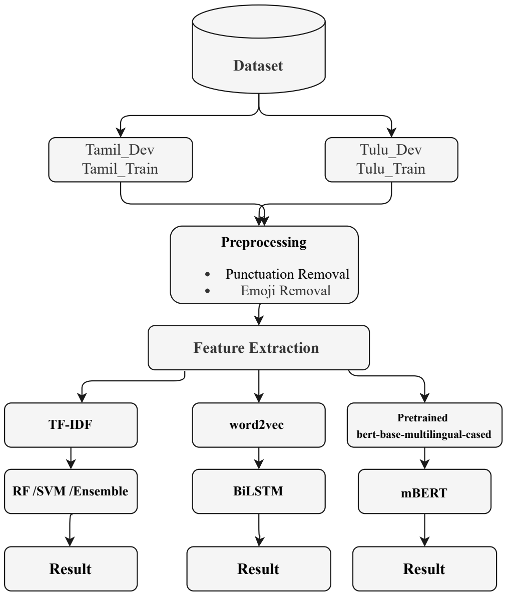

Figure 1: Proposed methodology

3.1 Preprocessing

- **Text Preprocessing:** To improve uniformity and reduce noise, the text data was preprocessed. Emoji's and punctuation were removed as they don't aid in classifying sentiment polarity. Unnecessary spaces were also removed for clean text input for further processing. To maintain the dataset's originality and the linguistic structure of code-mixed text, we did not use any extra preprocessing approaches in this work.

- **Label Encoder:** The sklearn's LabelEncoder is used to change text sentiment categories from Tamil and Tulu datasets into numbers. The Tulu dataset assigns numeric values for mixed feelings: negative, neutral, not Tulu and positive (0, 1, 2, 3, 4). The Tamil dataset uses mixed feelings, negative, positive, and unknown states coded as (0, 1, 2, 3). The data is transformed into numbers for consistent training and validation. The "inverse_transform()" method converts numeric labels back to their original categories, aiding in sentiment class conversion and evaluations among datasets.

3.2 Feature Extraction

In machine learning, Word2Vec and TF-IDF are frequently used for text representation. Scikit-learn's TF-IDF concentrates on word frequency while disregarding semantics, whereas Gensim's Word2Vec uses dense embeddings to capture semantic relationships.

- TF-IDF: Preprocessed text is transformed into numerical features using TF-IDF (Ahuja et al., 2019). A custom tokenizer processes Tamil and Tulu scripts. For sentiment classification in a code-mixed environment, the TF-IDFVectorizer selects the top 1000 features, addresses class imbalance, and reduces dimensionality.

- Word2Vec: Word2Vec creates dense vector embeddings based on contextual relationships (Jatnika et al., 2019). It captures syntactic and semantic relationships in code-mixed text and is used to generate embeddings stored in an embedding_matrix for deep learning models. Unlike TF-IDF, which focuses on word frequency, Word2Vec handles informal, multilingual social media text by encoding contextual meanings.

- Pretrained mBERT (bert-base-multilingual-cased): Trained on 104 languages, including Tamil and Tulu (Manias et al., 2023), mBERT is a transformer-based model that efficiently processes code-mixed text, maintaining case distinctions through WordPiece tokenization. It captures contextual meanings, improving sentiment classification in code-mixed Tamil-English and Tulu-English data. Deep contextual information is naturally captured by mBERT; therefore, more sophisticated feature extraction techniques like FastText or other transformer-based embeddings were not used separately.

4 Model Building

4.1 Random Forest

An ensemble learning model that aggregates several decision trees to improve the robustness of sentiment analysis. For code-mixed text, it enhances generalization. For assessing performance, GridSearchCV uses a five-fold cross-validation procedure to optimize hyperparameters, tuning estimators, tree depth, and feature selection.

4.2 Support Vector Machine

Using a linear kernel, the supervised learning algorithm divides sentiment into categories. TF-IDF uses bigrams and unigrams to preprocess, tokenize, and vectorize the text. F1-score, recall, accuracy, and precision are used to assess the model. LabelEncoder is used to accurately detect polarity by classifying sentiment predictions.

4.3 Ensemble

The stacking classifier improves accuracy by combining SVM and RF predictions. Whereas SVM determines the best hyperplane for emotion differentiation, RF uses decision trees to capture complex phenomena. In the final logistic regression model, a scaler standardizes predictions, utilizing the advantages of both models to improve sentiment classification.

4.4 BiLSTM

The BiLSTM architecture processes sequential data using both forward and backward context analysis (Wang, 2023). Tokenization comes first, then the creation of sequences and padding. With Word2Vec, word embeddings are produced while preserving semantic relationships. Pre-trained weights are used in the embedding layer of the BiLSTM model, which also has a softmax layer for sentiment classification and dense layers for feature extraction. It is trained with contextual learning and semantic representations using the Adam optimizer and categorical cross-entropy loss over 10 epochs with a batch size of 32.

4.5 mBERT

The mBERT reads Tamil-Tulu code-mixed text after being trained on 104 languages. Attention masks are applied to tokenized inputs (padded to 128) during processing. Sentiment is encoded with one hot (four classes for Tulu, six classes for Tamil). The model uses an AdamW optimizer, a learning rate scheduler, and categorical cross-entropy loss to train over two epochs. Contextual embeddings in mBERT improve sentiment analysis in social media content.

5 Experiment Analysis

5.1 Dataset

The data set for this study is from DravidianLangTech@NAACL2025's Shared Task on Sentiment Analysis in Tamil and Tulu (Durairaj et al., 2025) (Chakravarthi et al., 2020), (Hegde et al., 2022), (Hegde et al., 2023). It includes YouTube comments with sentiment labels for training, development, and testing. The task is to classify sentiments in code-mixed text into categories of Tulu and Tamil datasets. A brief description of the dataset is presented in Tables 1 and 2.

Dataset	Not Tulu	Positive	Neutral	Mixed	Negative
Tulu_Train	4,400	3,769	3,175	1,114	843
Tulu_Dev	543	470	368	143	118

Table 1: Description of Tulu Dataset

Dataset	Positive	Unknown State	Negative	Mixed Feelings
Tamil_Train	18,145	5,164	4,151	3,662
Tamil_Dev	2,272	619	480	472

Table 2: Description of Tamil Dataset

The Tamil and Tulu datasets show opportunities and challenges for sentiment analysis. The Tamil dataset has more samples, especially for positive sentiment, but both datasets are adequately sized. They have significant class imbalances, particularly in the Tulu dataset, which favors the Not Tulu category. Advanced emotional labels add complexity, and the mix of Tamil, Tulu, and English requires careful modeling. Despite challenges, these datasets can enhance sentiment analytics using data augmentation and deep learning methods.

5.2 Result Analysis

A detailed discussion of the evaluation metrics in this study is discussed in Tables 3 and Table 4. The performance of the model is assessed using four main metrics: F1-score, macro avg, weighted avg and accuracy.

Method	F1-Score	Macro Avg	Weighted Avg	Accuracy
BiLSTM	0.54	0.28	0.48	0.53
RF	0.63	0.41	0.60	0.62
Ensemble	0.63	0.41	0.60	0.62
SVM	0.65	0.43	0.62	0.64
mBERT	0.69	0.46	0.67	0.68

Table 3: Experimental Analysis using DravidianLangTech@NAACL Tulu datasets

Method	F1-Score	Macro Avg	Weighted Avg	Accuracy
BiLSTM	0.52	0.23	0.44	0.51
SVM	0.62	0.36	0.55	0.62
Ensemble	0.62	0.39	0.56	0.61
RF	0.61	0.39	0.56	0.62
mBERT	0.65	0.43	0.59	0.64

Table 4: Experimental Analysis using DravidianLangTech@NAACL Tamil datasets

Analysis of Tulu and Tamil datasets shows that mBERT is the top model, achieving the highest F1-Score and accuracy or both languages. For Tulu, mBERT reaches an F1-Score of 0.69 and an accuracy of 0.68, while for Tamil it scores 0.65 F1-Score and 0.64 accuracy. SVM shows slight improvements in Tulu (0.65 F1-Score) compared to Tamil (0.62 F1-Score), with the same precision of 0.64. Ensemble methods do not outperform individual models, and BiLSTM is the least effective in Tamil, with an F1-Score of 0.52. Overall, mBERT is the best choice, while BiLSTM is the weakest. Figure 2 presents a comparison of F1-Score between Tulu and Tamil.

Ensemble models exhibit promise for Tamil and Tulu datasets; however, there is a need for enhancement. The F1-score of mBERT decreased from 0.65 to 0.5511 on the Tulu test set, attributed to overfitting, while the F1-score of the ensemble model declined from 0.63 to 0.4117 on the Tamil test set, indicating issues with generalization. To improve performance, it is essential to increase the training data, implement dropout, apply L2 regularization, fine-tune hyperparameters, and utilize data augmentation. Additional improvements may be realized through semi-supervised learning, cross-validation, and sophisticated ensemble methods, especially for languages with limited resources.

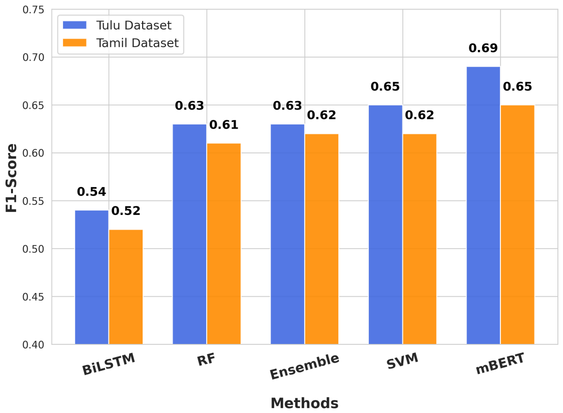

Figure 2: F1-Score Comparison: Tulu Vs Tamil

5.3 Limitations

The suggested approach is constrained by its reliance on pre-trained embeddings, which might not adequately capture domain-specific subtleties in some languages, despite its encouraging results. Additionally, noisy or inadequate data can cause performance to deteriorate, particularly for languages with limited resources like Tulu. The class imbalance in the dataset affects the performance of the model, resulting in biased predictions. In code-mixed text, the models also struggle to handle intricate linguistic structures.

6 Conclusion and Future Work

The research explores methods of machine learning and deep learning for the analysis of sentiments in Tamil and Tulu code-mixed data. We discovered that mBERT performed better than other models in terms of accuracy and F1-score while analyzing social media data, but SVM and ensemble methods performed well when dealing with unbalanced data. However, further normalizing is needed for transformer models such as mBERT in order to improve generality. In order to improve performance, future research can investigate SMOTE, back-translation, class-weighted loss, and semi-supervised learning. By concentrating on sentiment recognition, this study establishes the groundwork for extending natural language processing in Dravidian languages and enhancing tools for lesser-known languages to support reproducibility and further research.

6.1 Code Availability

The code for this study is available on GitHub - https://github.com/RakshaAdyanthayaA/Codalab-Shared-TAsk.

References

Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, and Pratyush Ahuja. 2019. The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, 152:341–348.

Faiza Mohammad Al-Kharboush and Mohammed Abdullah Al-Hagery. 2021. Features extraction effect on the accuracy of sentiment classification using ensemble models. *International Journal of Science and Research*, 10(3):228–231.

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.

Toqeer Ehsan, Amina Tehseen, Kengatharaiyer Sarveswaran, and Amjad Ali. 2023. Sentiment analysis of code-mixed tamil and tulu by training contextualized elmo representations. *RANLP'2023*, page 152.

M Ali Fauzi. 2018. Random forest approach fo sentiment analysis in indonesian. *Indones. J. Electr. Eng. Comput. Sci*, 12:46–50.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. Kancmd: Kannada codemixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63.

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, SK Lavanya, Durairaj Thenmozhi, 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.

Doaa Mohey El-Din Mohamed Hussein. 2018. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30(4):330–338.

Derry Jatnika, Moch Arif Bijaksana, and Arie Ardiyanti Suryani. 2019. Word2vec model analysis for semantic similarities in english words. *Procedia Computer Science*, 157:160–167.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

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.

George Manias, Argyro Mavrogiorgou, Athanasios Kiourtis, Chrysostomos Symvoulidis, and Dimosthenis Kyriazis. 2023. Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. *Neural Computing and Applications*, 35(29):21415–21431.

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*.

B Prathvi, K Manavi, K Subrahmanyapoojary, Asha Hegde, G Kavya, and Hosahalli Shashirekha. 2024. Mucs@ dravidianlangtech-2024: A grid search approach to explore sentiment analysis in code-mixed tamil and tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 257–261.

Wenliang Wang. 2023. Text sentiment classification method based on bilstm. *Highlights in Business, Economics and Management*, 21:679–687.