KEC_TECH_TITANS@DravidianLangTech 2025:Sentiment Analysis for Low-Resource Languages: Insights from Tamil and Tulu using Deep Learning and Machine Learning Models

Malliga Subramanian[1], Kogilavani S V[1], Dharshini S[1], Deepiga P[1], Praveenkumar C[1], Ananthakumar S[1]

[1]*Kongu Engineering College, Erode, Tamil Nadu, India*

Abstract

Sentiment analysis in Dravidian languages like Tamil and Tulu presents significant challenges due to their linguistic diversity and limited resources for natural language processing (NLP). This study explores sentiment classification for Tamil and Tulu, focusing on the complexities of handling both languages, which differ in script, grammar, and vocabulary. We employ a variety of machine learning and deep learning techniques, including traditional models like Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), as well as advanced transformer-based models like BERT and multilingual BERT (mBERT). A key focus of this research is to evaluate the performance of these models on sentiment analysis tasks, considering metrics such as accuracy, precision, recall, and F1-score. The results show that transformer-based models, particularly mBERT, significantly outperform traditional machine learning models in both Tamil and Tulu sentiment classification. This study underscores the need for further research to overcome challenges like language-specific nuances, dataset imbalance, and data augmentation techniques for improved sentiment analysis in under-resourced languages like Tamil and Tulu.

1 Introduction

Sentiment analysis in Tamil and Tulu is challenging due to their linguistic diversity and the lack of annotated datasets. Tamil is spoken in India and Sri Lanka, while Tulu is mainly used in coastal Karnataka and Kerala. Their unique linguistic features make sentiment analysis crucial for opinion mining, social media monitoring, and customer feedback analysis. The challenge increases with code-mixing, where users frequently switch between Tamil, Tulu, and English, especially on social media, making it harder to train and evaluate models effectively. Recent advancements in machine learning and deep learning have improved sentiment analysis, with transformer-based models like BERT and mBERT achieving state-of-the-art results. However, their application in Tamil and Tulu remains underexplored (Durairaj et al., 2025). This study evaluates traditional models such as Logistic Regression, SVM, and KNN, alongside advanced models like BERT and mBERT, for sentiment classification in Tamil and Tulu. We analyze how well they handle code-mixing and data scarcity, offering insights to improve sentiment analysis in low-resource languages and develop better NLP tools for Tamil and Tulu. Additionally, this research aims to bridge the gap in sentiment analysis for Dravidian languages by exploring more effective machine learning approaches.

2 Literature Survey

Sentiment analysis is a vital area of research in natural language processing (NLP), with significant applications in opinion mining, social media monitoring, and customer feedback analysis. While considerable progress has been made in sentiment classification for high-resource languages like English, research for low-resource languages, particularly Dravidian languages such as Tamil and Tulu, remains underexplored. The complexity of sentiment analysis in these languages arises from their diverse linguistic structures, code-switching, and cultural nuances, which make it a challenging task for machine learning and deep learning models. Additionally, the lack of large, annotated datasets in Tamil and Tulu further complicates the development of robust models for sentiment classification.

Thenmozhi et al. (2025) provide an extensive overview of sentiment analysis for Tamil and Tulu, detailing the challenges and methods used in previous research (Durairaj et al., 2025). Their study emphasizes the significance of transformer-based models, particularly multilingual BERT (mBERT),

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 356–360

in improving sentiment classification performance. The work also highlights the difficulties associated with code-mixing, data scarcity, and the necessity of annotated datasets for effective sentiment analysis in low-resource languages.

2.1 Sentiment Analysis in Tamil

Sentiment analysis in Tamil has evolved significantly over the years. Early studies relied on traditional machine learning models like Support Vector Machines (SVM) and Naive Bayes (NB), using features such as n-grams and sentiment lexicons. (Prabhu and Sundararajan, 2014) achieved moderate success with accuracy rates of 70-80%, but the use of manually created resources and limited ability to capture context restricted model performance.

In recent years, deep learning approaches have been adopted, offering improvements in understanding contextual meaning. (Babu and Ranjan, 2020) used Convolutional Neural Networks (CNN) for sentiment analysis in Tamil, yielding better results compared to traditional methods. More recently, (Ranjan and Babu, 2021) applied BERT-based models, which outperformed previous methods due to their ability to learn contextual embeddings, marking a significant advancement in Tamil sentiment analysis.

2.2 Sentiment Analysis in Tulu

Sentiment analysis in Tulu, a low-resource Dravidian language, has been underexplored due to limited linguistic resources and annotated corpora. Early attempts at Tulu sentiment classification were largely reliant on traditional machine learning models, such as Support Vector Machines (SVM) and Naive Bayes (NB), often utilizing handcrafted features like n-grams and lexicon-based approaches. These methods faced challenges due to the lack of large-scale labeled datasets, and their performance was constrained by the language's unique syntactic structure and informal usage in social media and online platforms. Additionally, rule-based approaches and domain-specific lexicons were employed, but they struggled to capture the complexities and nuances of Tulu sentiment expressions.

In recent years, there has been a shift towards leveraging deep learning models, particularly transformer-based architectures like multilingual BERT (mBERT), which are pre-trained on large multilingual corpora and can be fine-tuned for low-resource languages such as Tulu. These models, with their capacity to learn contextual embeddings and capture long-range dependencies, have demonstrated superior performance in sentiment analysis for other Dravidian languages, such as Tamil and Malayalam. Transfer learning, where models trained on high-resource languages are adapted to Tulu, is emerging as a promising strategy to overcome data limitations. However, the scarcity of annotated Tulu datasets remains a significant challenge, underscoring the need for further research and the development of comprehensive labeled corpora to improve the robustness and accuracy of sentiment classification systems for Tulu.

2.3 Challenges in Sentiment Analysis for Dravidian Languages

Several challenges hinder the progress of sentiment analysis in Dravidian languages. One of the key difficulties is the lack of large, annotated datasets, which are essential for training robust models. Dravidian languages have diverse morphological structures, dialects, and variations in colloquial language, which further complicates sentiment detection.

Another significant challenge is code-switching, where speakers alternate between Dravidian languages and languages like English, especially in digital communication. (Subashini et al., 2022) found that sentiment analysis in code-mixed data leads to a decline in model performance, as models trained on single-language datasets struggle to interpret the multilingual input effectively.

2.4 Recent Advances and Transformer Models

Recent advancements in NLP, particularly with transformer-based models such as BERT and its multilingual variant mBERT, have shown promising results in sentiment analysis tasks for low-resource languages. (Vaswani et al., 2017) introduced the Transformer architecture, which revolutionized NLP tasks by providing a method to capture long-range dependencies in text without relying on sequential processing, as was the case with RNNs and LSTMs.

In the context of Dravidian languages, (Ghosal et al., 2020) demonstrated the utility of mBERT in multilingual sentiment analysis tasks, where it outperformed traditional machine learning approaches by leveraging cross-lingual transfer learning. This has opened new avenues for handling sentiment analysis in under-resourced languages like Tamil

and Tulu, as mBERT is pre-trained on a large corpus of multilingual data, providing a foundation to fine-tune models for specific languages or tasks.

3 Materials and Methods

3.1 Dataset Description

The dataset used in this study consists of Tamil-Tulu code-mixed text collected from social media platforms such as Twitter, Facebook, and regional online forums. It contains 8,000 training samples and 2,000 validation samples. The sentiment class distribution in the Tamil dataset is as follows: 58.30% Positive, 13.34% Negative, and 11.77% Mixed Feelings, with an additional 16.59% categorized as "Unknown State" that required filtering. In the validation set, the distribution remains similar, with 59.12% Positive, 12.49% Negative, and 12.28% Mixed Feelings, alongside 16.11% Unknown State.

For the Tulu dataset, a significant proportion (33.08% in training and 33.07% in validation) consists of "Not Tulu" samples, which were excluded from sentiment classification. Among the actual sentiment labels, 28.34% were Positive, 23.87% Neutral, 8.38% Mixed, and 6.34% Negative in the training set, while the validation set showed 28.62% Positive, 22.41% Neutral, 8.71% Mixed, and 7.19% Negative. The dataset also exhibited class imbalance, particularly in the Mixed category, which was addressed using oversampling techniques like SMOTE and random resampling.

3.2 Pre-processing and Feature Extraction

Preprocessing and feature extraction for sentiment analysis in Tamil and Tulu involve several key steps to handle the unique linguistic challenges of these languages. Text normalization standardizes the text by lowercasing, removing extra spaces, and handling informal contractions, while tokenization splits the text into words or subwords, especially in code-mixed contexts.

Stopwords and noise, such as emojis, special characters, and URLs, are removed to reduce irrelevant information. Lemmatization or stemming is applied to reduce words to their base forms, and code-switching between Tamil, Tulu, and English is handled using language identification and transliteration techniques. Feature extraction includes techniques like Bag-of-Words (BoW) and TF-IDF to capture word frequencies, as well as more advanced approaches like word embeddings (Word2Vec, FastText) and transformer-based models (BERT, mBERT) to capture semantic meaning and contextual nuances. Additionally, sentiment lexicons specific to Tamil and Tulu can be integrated to enhance the detection of sentiment-bearing words. These preprocessing and feature extraction steps enable effective sentiment classification by preparing the data for machine learning and deep learning models.

3.3 Proposed Classifiers

To classify sentiment in Tamil and Tulu, we used both traditional machine learning and deep learning models. SVM and KNN were chosen for their efficiency with small datasets and clear decision boundaries, while Decision Trees helped explore rule-based classification but struggled with complex language structures and code-mixing.

Deep learning models like CNNs and RNNs captured patterns and sequential dependencies, making them better suited for morphologically rich languages. However, transformer-based models like BERT and mBERT performed best due to their strong contextual understanding and multilingual capabilities. Given the frequent code-mixing in Tamil-Tulu, mBERT excelled due to its pre-training on diverse languages. This combination of models helped balance context, sequence learning, and feature-based classification to improve sentiment analysis accuracy.

4 Results and Discussion

The sentiment analysis results for Tamil and Tulu indicate that transformer-based models like BERT and mBERT delivered the highest accuracy due to their ability to capture complex contextual information in code-mixed text. CNNs effectively detected local patterns, while RNNs and GRUs excelled in handling sequential dependencies across sentences. MLPs provided reasonable performance, while Random Forests offered robustness but slightly lower accuracy. Logistic Regression, though efficient for binary classification, performed less effectively than deep learning models. Combining these approaches enhanced overall accuracy.

4.1 Performance Metrics

The performance metrics for sentiment analysis in Tamil and Tulu include accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC curve. Accuracy measures the overall proportion of correct classifications, while precision evaluates the

ability to avoid false positives, and recall assesses the model's capability to capture all relevant sentiment instances. The F1-score balances precision and recall, providing a comprehensive measure. The confusion matrix breaks down classification errors, and the AUC-ROC curve indicates the model's ability to distinguish between different sentiment classes. These metrics offer a thorough evaluation of the model's effectiveness, especially in the context of code-mixed text and language-specific nuances.

Table 1: Performance Table of our Models for Sentiment Analysis in Tamil and Tulu

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
KNN	56	64	63	64
SVM	77	83	80	83
DT	73	68	65	66
BERT	85	80	78	82
RNN	70	65	60	63

4.2 Limitations

While our approach improves sentiment classification for Tamil and Tulu, several challenges remain.

First, dataset imbalance affects performance, especially in the Mixed Feelings category. Although oversampling helped, advanced data augmentation techniques could further improve balance. Second, high computational costs of deep learning models like BERT and mBERT make real-time sentiment analysis challenging. Using lighter transformer models or compression techniques could address this issue.

Another challenge is code-mixed text, where Tamil, Tulu, and English are mixed in a single sentence. While mBERT performed well, it was not trained specifically for Tamil-Tulu mixing, causing occasional errors. Future work could involve fine-tuning transformers on Tamil-Tulu datasets. Finally, limited annotated data affects accuracy. Expanding the dataset with more labeled examples and domain-specific lexicons would improve performance.

5 Conclusion and Future Work

In conclusion, this study highlighted the effectiveness of machine learning and deep learning models, particularly transformer-based models like BERT and mBERT, for sentiment analysis in Tamil and Tulu. These models outperformed traditional methods by capturing contextual nuances in code-mixed text. While CNNs, RNNs, and GRUs showed strong performance in identifying patterns and sequential dependencies, simpler models like Logistic Regression were less effective. Future work can focus on expanding the dataset, addressing class imbalance, integrating domain-specific lexicons, and fine-tuning multilingual models like mBERT to further improve sentiment analysis accuracy and model generalization for Tamil and Tulu.

Project Repository

The full source code for this project is available on GitHub: GitHub Repository Deepikagowtham

References

D. Babu and S. Ranjan. 2020. Sentiment Analysis of Tamil Text Using Convolutional Neural Networks. In *Proceedings of the 5th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2020)*.

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.

D. Ghosal, M. Thakur, and A. Patel. 2020. Multilingual BERT for Sentiment Analysis in Low-Resource Dravidian Languages. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*.

A. Prabhu and V. Sundararajan. 2014. Sentiment Classification for Tamil Language Using Machine Learning Algorithms. *International Journal of Computer Applications*, 97(5):28–32.

S. Ranjan and D. Babu. 2021. Exploring BERT for Sentiment Analysis in Tamil: A Deep Learning Approach. In *Proceedings of the 6th International Conference on Artificial Intelligence and Data Science (AIDS 2021)*.

R. Subashini, P. Kumar, and L. Devi. 2022. Challenges in Sentiment Analysis of Code-Mixed Dravidian Languages. *International Journal of Computational Linguistics*, 14(2):102–118.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 5998–6008.