RMKMavericks@DravidianLangTech 2025: Emotion Mining in Tamil and Tulu Code-Mixed Text: Challenges and Insights

Gladiss Merlin N.R
R.M.K. Engineering College
Tiruvallur
nrg.ad@rmkec.ac.in

Boomika E
R.M.K. Engineering College
Tiruvallur
boom22011.ad@rmkec.ac.in

Lahari P
R.M.K. Engineering College
Tiruvallur
laha22024.ad@rmkec.ac.in

Abstract

Sentiment analysis in code-mixed social media comments written in Tamil and Tulu presents unique challenges due to grammatical inconsistencies, code-switching, and the use of non-native scripts. To address these complexities, we employ pre-processing techniques for text cleaning and evaluate machine learning models tailored for sentiment detection. Traditional machine learning methods combined with feature extraction strategies, such as TF-IDF, are utilized. While logistic regression demonstrated reasonable performance on the Tamil dataset, achieving a macro F1 score of 0.44, support vector machines (SVM) outperformed logistic regression on the Tulu dataset with a macro F1 score of 0.54. These results demonstrate the effectiveness of traditional approaches, particularly SVM, in handling low-resource, multilingual data, while also highlighting the need for further refinement to improve performance across underrepresented sentiment classes.

1 Introduction

The growing use of social media platforms has led to an abundance of user-generated content, often expressed in code-mixed languages. Tamil and Tulu, two Dravidian languages, frequently appear in such code-mixed forms, blending with English and other languages. These multilingual and code-mixed texts introduce significant challenges for sentiment analysis due to their informal grammar, irregular structures, and non-standard scripts.

Sentiment analysis aims to identify and classify subjective opinions or emotions expressed in text. While considerable progress has been made in analyzing texts in resource-rich languages, low-resource languages like Tamil and Tulu remain underexplored. Existing tools and models, primarily designed for monolingual texts, struggle to perform effectively on code-mixed data, necessitating novel approaches tailored to these contexts.

In this work, we investigate traditional machine learning methods for sentiment analysis on Tamil-English and Tulu-English code-mixed datasets. We leverage pre-processing techniques to clean and normalize the data, employ TF-IDF (Term Frequency-Inverse Document Frequency) for feature extraction, and evaluate multiple machine learning classifiers. By optimizing hyperparameters and focusing on feature engineering, we aim to improve sentiment classification performance for these low-resource, multilingual datasets. The results underscore the importance of adapting traditional techniques to the unique challenges posed by code-mixed text data.

2 Related Work

Sentiment analysis has been a prominent area of research for several decades, focusing on identifying and classifying emotions, opinions, and attitudes expressed in text. Traditional approaches to sentiment analysis can be categorized into lexicon-based, machine learning-based, and hybrid methods. Lexicon-based approaches rely on predefined sentiment dictionaries to determine the polarity of text. Machine learning-based methods, often using supervised algorithms, leverage labeled datasets to train models for sentiment classification. Hybrid methods combine the strengths of both approaches, aiming to improve accuracy across diverse datasets.

Research in sentiment analysis for low-resource languages, such as Tamil and Tulu, has gained attention more recently. Early studies predominantly relied on rule-based systems or basic machine learning techniques, often limited by the availability of annotated datasets and language-specific tools. For example, Thavareesan and Mahesan (2020a) explored machine learning techniques for Tamil text sentiment analysis using feature representations like word embeddings and TF-IDF.

https://github.com/Boomika2005/RMKMavericks-Sentiment-analysis

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 144–147
May 3, 2025 ©2025 Association for Computational Linguistics

Sentiment analysis of code-mixed text introduces additional complexities due to frequent switching between languages, irregular grammar, and the use of non-native scripts.

Recent studies have also examined the use of deep learning models, such as recurrent neural networks and transformer architectures, for low-resource and code-mixed languages. However, these methods typically require substantial computational resources and large annotated datasets, which are not always available for Tamil and Tulu.

Building on this body of work, our research focuses on traditional machine learning models optimized with feature extraction techniques, such as TF-IDF, and hyperparameter tuning. By leveraging these methods, we aim to address the unique challenges posed by Tamil-English and Tulu-English code-mixed datasets.

3 Task details

Sentiment analysis refers to the process of determining the emotional tone or subjective opinions expressed in a given piece of text. This area of research has gained significant attention over the past two decades, both in academic and industry settings. With the rise of social media, there is an increasing demand for systems that can analyze sentiment in social media posts, which are often written in code-mixed languages, particularly in Dravidian languages. Code-mixing, the practice of blending multiple languages in a single sentence or passage, is common in multilingual communities, and these texts may sometimes be written in non-native scripts. Traditional systems trained on monolingual datasets struggle with code-mixed text due to the complexities of language switching and its varying impact on grammar, syntax, and vocabulary.

The objective of this task is to determine the sentiment polarity of code-mixed comments/posts in Tamil-English and Tulu-English, collected from social media platforms. While each comment/post may consist of more than one sentence, the average sentence length in this dataset is short. Each comment/post is labeled with a sentiment polarity, either positive, negative, or neutral. The dataset also includes class imbalance, reflecting the real-world challenges encountered in sentiment analysis applications.

This task encourages further exploration into how sentiment is expressed in code-mixed texts, particularly in the context of social media communications.

4 Methodology

Our approach for sentiment analysis in the shared task involved implementing two traditional machine learning models and performing various data processing steps to handle the challenges of code-mixed text. We began by importing essential libraries such as Pandas, NumPy, and scikit-learn for tasks like data loading, cleaning, tokenization, vectorization, and modeling.

First, we loaded the training and validation datasets using Pandas, which contained code-mixed Tamil-English and Tulu-English comments/posts. We then cleaned the data by removing unnecessary punctuation and converting the text to lowercase to ensure consistency. This preprocessing step helped improve the models' ability to detect sentiment accurately.

After cleaning the text, we applied the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer to transform the text data into a numerical format that can be used by machine learning models. We selected unigrams and bigrams for tokenizing the text, capturing both individual words and word pairs to retain the context of code-switching in the text.

We trained two machine learning models: Logistic Regression and Support Vector Machine (SVM). The logistic regression model was trained using the 'liblinear' solver, suitable for small datasets, while the SVM model was trained with different kernel functions, including linear and radial basis function (RBF), along with a grid search for tuning hyperparameters such as C and gamma.

To address class imbalance in the dataset, we used techniques such as adjusting class weights during model training to ensure the model pays appropriate attention to underrepresented classes. Hyperparameter tuning was performed using GridSearchCV to identify the optimal parameters for each model and improve performance.

Once the models were trained, we evaluated their performance on the validation dataset using evaluation metrics such as accuracy and the classification report. The classification report provided detailed insights into the model's precision, recall, and F1-score for each sentiment class.

In the final step, we selected the best-performing model based on the validation performance and saved the model, the TF-IDF vectorizer, and the label encoder using joblib. These components can be loaded later for deployment or future predictions on unseen data.

Our methodology provided a robust framework for sentiment analysis on code-mixed text, leveraging traditional machine learning models and effective text preprocessing techniques to tackle the complexities of code-switching in social media comments.

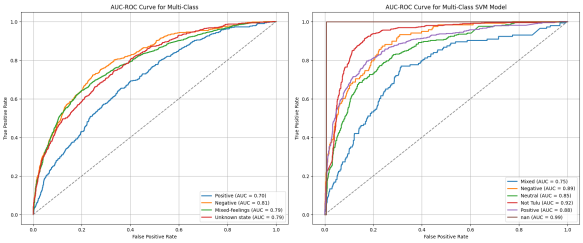

Figure 1: AUC-ROC Curve for Tamil & Tulu

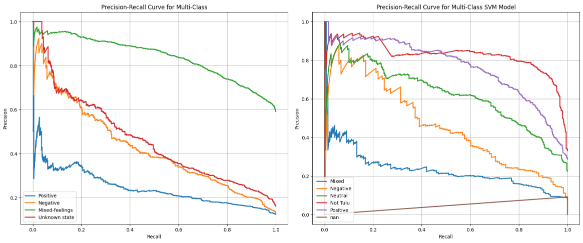

Figure 2: Precision-Recall for Tamil & Tulu

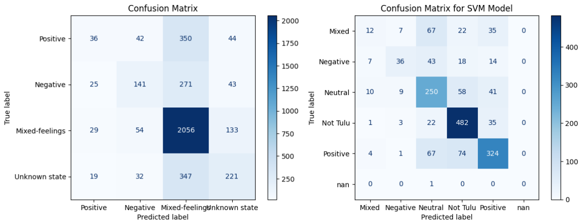

Figure 3: Confusion Matrix for Tamil & Tulu

5 Result and Findings

In our evaluation of machine learning models for sentiment analysis, we utilized several performance metrics, including Accuracy, Precision, Recall, and Macro F1 Score. The experiments involved two models: Logistic Regression (LR) and Support Vector Machine (SVM), both using TF-IDF feature extraction. Overall, the experiments demonstrate the effectiveness of traditional machine learning approaches for sentiment analysis in low-resource, code-mixed datasets while highlighting areas for improvement to handle underrepresented sentiment classes better.

Metrics	**Tamil**	**Tulu**
Accuracy	0.44	0.54
Precision	0.44	0.54
Recall	0.44	0.54
Macro F1 Score	0.4354	0.5318

Table 1: Tamil and Tulu Classification Report.

6 Conclusion

This study explored sentiment analysis on code-mixed Tamil-English and Tulu-English social media text using machine learning models, specifically Logistic Regression and Support Vector Machine (SVM). The results demonstrated that Logistic Regression outperformed SVM, achieving higher macro F1 scores and showing a better ability to detect sentiment polarity. The TF-IDF feature extraction method played a significant role in capturing the essential features from the code-mixed text. Although the models performed well overall, challenges such as class imbalance were observed, affecting the classification of minority sentiment classes. Adjusting class weights helped alleviate some of these issues. Future work could involve enhancing model performance with more advanced approaches, such as deep learning techniques (e.g., LSTM or BERT), to better address the complexities of code-switching. Overall, this study underscores the potential of machine learning for sentiment analysis in code-mixed social media data, while highlighting opportunities for further refinement and optimization.

7 Limitations

Despite the encouraging outcomes of our method in sentiment analysis of code-mixed Tamil-English and Tulu-English text, several difficulties still exist. Among the main drawbacks is the dataset's class imbalance, which impairs the model's capacity to correctly categorize sentiment classes that are underrepresented. The imbalance affected overall performance even after class weight adjustments somewhat alleviated this problem. Additional challenges for conventional machine learning models were the intricacy of code-mixed text, which included non-standard scripts, frequent language change, and irregular grammar. Even though TF-IDF-based feature extraction worked well, it might not adequately capture words' contextual meaning in mixed-language constructions. In the future, it could be possible to incorporate more sophisticated methods that preserve computing efficiency while better understanding the subtleties of code-mixed text. Notwithstanding these difficulties, the study shows how machine learning may be used for sentiment analysis in multilingual, low-resource environments and lays the groundwork for further investigation and improvement.

References

Abdullah Alsaeedi and Mohammad Zubair Khan. 2019. Study on sentiment analysis techniques of twitter data. *International Journal of Advanced Computer Science and Applications*.

N.S. Athindran, S. Manikandaraj, and R. Kamaleshwar. 2018. Comparative analysis of customer sentiments on competing brands using hybrid model approach. In *Proceedings of the 2018 IEEE 3rd International Conference on Inventive Computation Technologies (ICICT)*, pages 348–353.

P. Chakriswaran, D.R. Vincent, K. Srinivasan, V. Sharma, C.Y. Chang, and D.G. Reina. 2019. Emotion ai-driven sentiment analysis: A survey, future research directions, and open issues. *Applied Sciences*, 9(5462).

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.

N. Iqbal, A.M. Chowdhury, and T. Ahsan. 2018. Enhancing the performance of sentiment analysis by using different feature combinations. In *Proceedings of the 2018 IEEE International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pages 1–4.

Y.G. Jung, K.T. Kim, B. Lee, and H.Y. Youn. 2016. Enhanced naive bayes classifier for real-time sentiment analysis with sparkr. In *Proceedings of the 2016 IEEE International Conference on Information and Communication Technology Convergence (ICTC)*, pages 141–146.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1–7.

Nishit Shrestha and Fatma Nasoz. 2019. Deep learning sentiment analysis of amazon.com reviews and ratings. *International Journal of Soft Computing and Artificial Intelligence (IJSCAI)*.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020. Sentiment lexicon expansion using word2vec and fasttext for sentiment prediction in tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

S. Vanaja and M. Belwal. 2018. Aspect-level sentiment analysis on e-commerce data. In *Proceedings of the 2018 IEEE International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1275–1279.

G. Vinodhini and R.M. Chandrasekaram. 2012. Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6):28–35.