codecrackers@DravidianLangTech 2025: Sentiment Classification in Tamil and Tulu Code-Mixed Social Media Text Using Machine Learning

Lalith Kishore V P [1], Manikandan G [1], Mohan Raj M A[1]
Keerthi Vasan A[1], Aravindh M [1]
[1]R.M.K. Engineering College, Tiruvallur, Tamilnadu, India
{lali22025, gmk, moha22029, keer22061, arav22001}.ad@rmkec.ac.in

Abstract

Sentiment analysis of code-mixed Dravidian languages has become a major area of concern with increasing volumes of multilingual and code-mixed information across social media. This paper presents the "Seventh Shared Task on Sentiment Analysis in Code-mixed Tamil and Tulu", which was held as part of DravidianLangTech(NAACL-2025). However, sentiment analysis for code-mixed Dravidian languages has received little attention due to issues with class imbalance, small sample size, and the very informal nature of the code-mixed text. This study applied an SVM-based approach for the sentiment classification of both Tamil and Tulu languages. The SVM model achieved competitive macro-average F1 scores of 0.54 for Tulu and 0.438 for Tamil, showing that traditional machine learning methods can address the problem of sentiment categorization in code-mixed languages under low-resource settings.

1 Introduction

Sentiment analysis consists of classifying text depending on the opinions and emotions of the writer expressed inside it. The DravidianLangTech shared task is meant to further research in this field by concentrating on code-mixed datasets of comments and posts, especially in low-resource languages such as Tamil and Tulu (S. K. et al., 2024). This task is a need of the day when social media platforms such as Instagram, X and YouTube serve as the principal portals for communication across the world, being used by a mass of people to express their opinions and emotions in a cross-linguistic or cross-geographical manner. This shared task helps develop robust sentiment classification models and establishes a benchmark for evaluating techniques in linguistically diverse settings. It addresses challenges in informal and code-mixed language use, contributing to more inclusive and effective sentiment analysis for low-resource languages. The rest of the paper is organized as follows. Section 2 outlines the related works emphasizing Sentiment Analysis in Dravidian languages. Section 3 presents a description of the dataset and data processing. Section 4 describes the methodology used for the shared task. Section 5 discusses the result and findings of the task assigned. In Section 6 concludes the paper. Section 7 highlights the limitations of this study. At last, we have the references.

2 Related Work

Sentiment analysis (SA) in code-mixed Dravidian languages has gained momentum with the rise of multilingual social media content. Foundational datasets like (Chakravarthi et al., 2020) for Tamil and (Hegde et al., 2022) for Tulu have enabled systematic exploration of code-mixed SA, though challenges persist in class imbalance, informal text, and low-resource settings. Several ML models are experimented with various features for SA of user-generated content in code-mixed low-resource languages (Hegde et al., 2023). Initial studies used standard machine learning (ML) models and applied feature extraction techniques like TF-IDF and Bag-of-Words (BoW). (Shanmugavadivel et al., 2024a) investigated Decision Trees, along with SVM, using Tamil code-mixed data; results showed high accuracy (99%) but low macro F1-scores (0.39), indicating a class imbalance problem. Like (B et al., 2024) a large ensemble of machine learning models with careful optimization was used, achieving considerably better macro F1-scores of 0.26 (Tamil) and 0.55 (Tulu). However, a closer look at the confusion matrices showed difficulties in distinguishing between subtle sentiments like "Mixed Feelings" and others. Several conventional machine learning approaches exhibit mean-

https://github.com/VPLALITHKISHORE/DravidianLangTech_SharedTask

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 465–469
May 3, 2025 ©2025 Association for Computational Linguistics

ingful limitations when confronted with the informal structure and skewed data distribution characteristic of code-mixed text, as these studies clearly show. Bi-LSTM and transformer architectures address contextual nuances. Roy and Kumar (2021) combined GloVe embeddings with Bi-LSTM for Tamil, achieving a weighted F1-score of 0.552 but faltering on minority classes. (Tripty et al., 2024) leveraged XLM-RoBERTa for Tulu (F1: 0.468) and used back-translation for Tamil, underscoring transformers potential despite data limitations. Class imbalance remains critical. (Kanta, 2023) observed F1-scores as low as 0.147 for Tamil using SVM, while (Shanmugavadivel et al., 2024a) noted Decision Trees 99% accuracy but 0.39 F1 due to skewed distributions. Recent solutions include data augmentation (Tripty et al., 2024) and hard-voting ensembles (B et al., 2024). (Ponnusamy et al., 2023) proposed ML models (LR, Multinomial Naive Bayes (MNB), and LinearSVC) trained with Term Frequency-Inverse Document Frequency (TF-IDF) of word unigrams for SA in Tamil and Tulu languages. Their proposed LR, MNB, and LinearSVC models obtained macro F1 scores of 0.43, 0.20, 0.41 and 0.51, 0.25, 0.49 for Tamil and Tulu languages respectively.

3 Dataset resource and data processing

To analyze sentiment in code-mixed languages, we leveraged existing datasets curated for sentiment analysis such as (Chakravarthi et al., 2020) for Tamil and (Hegde et al., 2022) for Tulu.

Labels	Train Set	Development Set	Test Set
Positive	18145	2272	1983
unknown_state	5164	619	593
Negative	4151	480	458
Mixed_feelings	3662	472	425
Total	31122	3843	3459

Table 1: Label-wise Breakdown of Tamil Code-Mixed Data.

The preprocessing phase involved standardizing and cleaning text data for both Tulu and Tamil languages to ensure consistency and reduce noise. While maintaining Tulu script characters using their Unicode range, text normalisation for Tulu involved changing all characters to lowercase and removing non-alphanumeric symbols. Similar normalisation was applied to Tamil text, with particular focus on keeping the characters from the Tamil script. Stopwords were eliminated by combining lists of English and language specific stopwords. Tamil used a comprehensive predefined list of stopwords, whereas Tulu used a custom-curated list. Using TF-IDF vectorization, feature extraction was carried out. Tamil employed unigrams just for simplicity, whereas Tulu used bigrams and unigrams to capture contextual subtleties. To ensure linguistic integrity, script-specific characters were kept intact throughout the tokenization process.

Labels	Train Set	Development Set	Test Set
Not Tulu	4400	543	474
Positive	3769	470	453
Neutral	3175	368	343
Mixed	1114	143	120
Negative	843	118	88
Total	13301	1642	1478

Table 2: Label-wise Breakdown of Tulu Code-Mixed Data.

4 Methodology

The methodology applied supports vector machines (SVMs), which were selected for this specific application because of their reliability with high-dimensional text data. In Tulu, class imbalance was countered with the generation of synthetic examples of the minority class via SMOTE oversampling, while hyper-parameter tuning through grid-search was applied to find optimal kernel and regularization parameters for the model. SMOTE was chosen over ADASYN because it generates synthetic samples evenly across the minority class, ensuring balanced augmentation without amplifying noise. ADASYN, which focuses on harder-to-learn examples, can introduce unwanted noise and overfitting, especially with high-dimensional TF-IDF features. SMOTE provides better stability for text classification. Conversely, the Tamil model used Class Weight Adjustment. Both models transformed text to TF-IDF vectors, Tulu's vectorizer being more oriented toward n-gram associations. The validation measures used included accuracy, precision, recall, and F1-scores-with Tulu's macro-averaged AUC-ROC being applied to assess multi-class performance. Classification reports contained performance details, sentiment distribution maps illustrated class balances, and confusion matrices revealed the patterns behind misclassifications; hence reproducibility and deployment readiness were provided by applying the optimized models in making the final predictions upon the test data with results stored for both languages.

4.1 Feature Extraction

The Tulu and Tamil datasets feature extraction procedure focused on the most informative words for sentiment classification by converting raw text into numerical vectors using the TF-IDF method. In order to capture both individual words and contextual phrases that may be essential for sentiment expression, such as negations or emotive combinations, both unigrams and bigrams (ngram_range=(1, 2)) were employed for the Tulu dataset. The feature space was limited to the top 5,000 most frequent terms (max_features=5000), allowing for efficient computation while retaining the most significant features. Furthermore, rare terms (those that appeared in fewer than three documents) were filtered out for the Tulu dataset using min_df=3, helping to eliminate noise.

4.2 Models and Techniques Utilized

For the Tulu task, an SVM model with GridSearchCV-optimized hyperparameters (kernel type, regularization C) was implemented. Class imbalance was addressed via SMOTE (synthetic oversampling). Features were extracted using TF-IDF (unigrams/bigrams, min_df=3) to filter rare terms. Performance was evaluated via macro-F1 (handling imbalance) and AUC-ROC (multi-class). For the Tamil task, a Linear SVM (C=1) with class weight adjustment was used to handle class imbalance instead of oversampling. TF-IDF focused on codemixed tokens (Tamil-English keywords). Metrics included accuracy and weighted F1 to assess sentiment across imbalanced classes. This can be observed with the help of visuals as in Table 3.

Component	Tulu	Tamil
Feature Extraction	Unigrams+bigrams, min-df=3	Unigrams only, codemix-aware tokens
Model	GridSearch-optimized	Linear SVM(fixed C=1)
Class Balancing	SMOTE oversampling	Class Weight Adjustment

Table 3: Label-wise Breakdown of Tulu and Tamil Code-Mixed Data

4.3 Classification for Tamil and Tulu codemix

The ability of the Tamil model to differentiate across various labels is demonstrated by the matrix. This can be observed with the help of visuals as in Figure 1. The ability of the Tulu model to differentiate across various labels is demonstrated by the matrix. This can be observed with the help of visuals as in Figure 2.

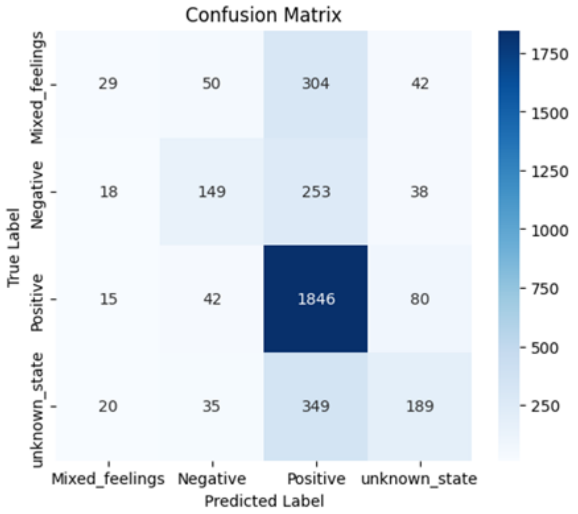

Figure 1: Confusion matrix of the proposed model for code-mixed Tamil text

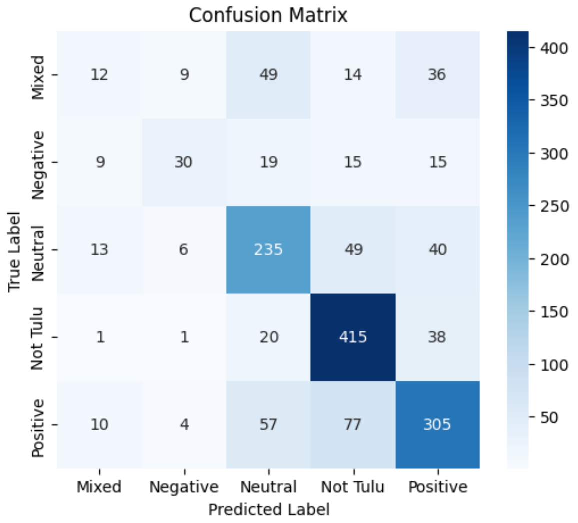

Figure 2: Confusion matrix of the proposed model for code-mixed Tulu text

5 Result and Findings

The performance of the model was evaluated across the labels based on the task. The models outperformed with this approach, achieving competitive macro average F1-scores of **0.54** for Tulu and **0.438** for Tamil. The sentiment distribution comparison of the labels and the classification report for Tamil can be observed with the help of visuals in Figure 3.

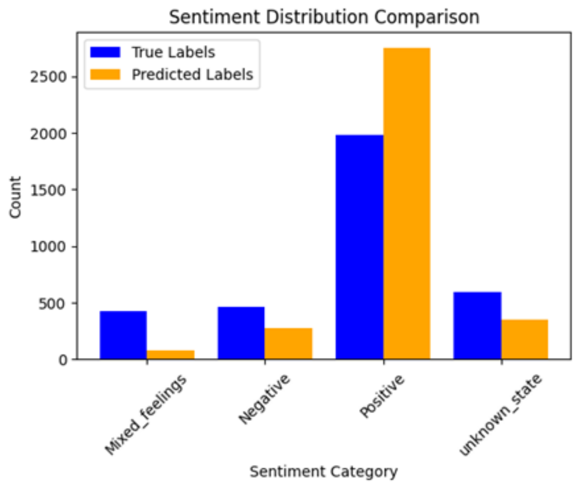

Figure 3: Sentiment Distribution Comparison and Classification report table for code-mixed Tamil.

The sentiment distribution comparison of the labels and the classification report for Tulu can be observed with the help of visuals in Figure 4.

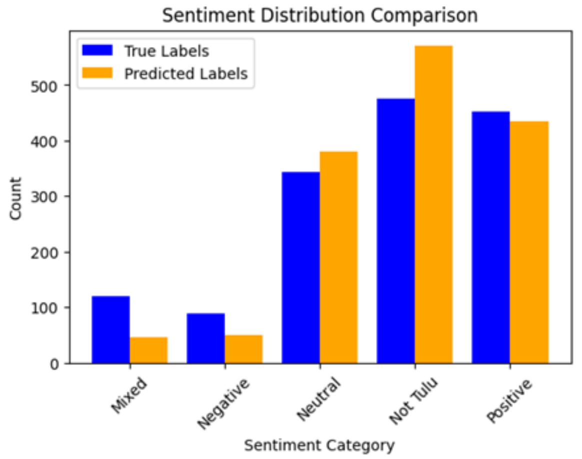

Figure 4: Sentiment Distribution Comparison and Classification report table for code-mixed Tulu.

6 Conclusion

SVM has proven to be a robust and adaptable approach for NLP tasks, particularly in low-resource settings. By leveraging hyperparameter tuning, SMOTE, and TF-IDF with n-grams, the model effectively handled class imbalance and noise in the Tulu dataset, achieving strong macro-F1 and AUC-ROC scores. For Tamil codemixed data, a Linear SVM with class weight adjustments and TF-IDF-based feature selection provided stable performance, as reflected in the weighted F1 score and accuracy. These results highlight the importance of tailoring preprocessing and optimization strategies to dataset characteristics—extensive balancing techniques for highly skewed distributions, and minimal interventions for well-represented codemixed contexts. Overall, SVM remains a powerful choice for sentiment classification, demonstrating its effectiveness in diverse linguistic and data imbalance scenarios.

7 Limitations

Tamil and Tulu SVM-based sentiment classification models face several limitations. The Tamil model, with a fixed linear SVM and class weight adjustment, may suffer from underfitting and struggles with colloquial expressions, transliterated words (Tanglish) and phrase-level sentiment detection due to its unigram-based TF-IDF approach. Although it partially handles Tanglish, it fails to capture complex code-mixed structures, sentiment shifts, and negations effectively. The Tulu model, while using SMOTE for imbalance correction, may introduce synthetic noise and relies on a manual stopword list, which might not fully cover dialectal variations.

References

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.

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.

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.

Lavanya S K, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Durairaj Thenmozhi, 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.

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.

Kogilavani Shanmugavadivel, Sowbharanika Janani J S, Navbila K, and Malliga Subramanian. 2024a. Code Maker@DravidianLangTech-EACL 2024: Sentiment Analysis in Code-Mixed Tamil using Machine Learning Techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Prathvi B, Manavi K K, Subrahmanya, Asha Hegde, Kavya G, and H L Shashirekha. 2024. MUCS@DravidianLangTech-2024: A Grid Search Approach to Explore Sentiment Analysis in Codemixed 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.

Pradeep Kumar Roy, and Abhinav Kumar. "Sentiment Analysis on Tamil Code-Mixed Text using Bi-LSTM." FIRE (Working Notes). 2021.

Selam Abitte 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*, Varna, Bulgaria.

Zannatul Fardaush Tripty, Md. Arian Al Nafis, Antu Chowdhury, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUETSentimentSillies@DravidianLangTechEACL2024: Transformer-based Approach for Sentiment Analysis in Tamil and Tulu Code-Mixed Texts. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Kishore Kumar Ponnusamy, Charmathi Rajkumar, Prasanna Kumar Kumaresan, Elizabeth Sherly, and Ruba Priyadharshini. 2023. VEL@ DravidianLangTech: Sentiment Analysis of Tamil and Tulu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 211–216. Varna, Bulgaria. Recent Advances in Natural Language Processing.