Code_Conquerors@DravidianLangTech 2025: Deep Learning Approach for Sentiment Analysis in Tamil and Tulu

Harish Vijay V, Ippatapu Venkata Srichandra, Pathange Omkareshwara Rao, Premjith B.
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham India
harishvijay0204@gmail.com, ippatapuvenkatasrichandra@gmail.com,
cb.en.u4aie22039@cb.students.amrita.edu, b_premjith@cb.amrita.edu

Abstract

In this paper we propose a novel approach to sentiment analysis in languages with mixed Dravidian codes, specifically Tamil-English and Tulu-English social media text. We introduce an innovative hybrid deep learning architecture that uniquely combines convolutional and recurrent neural networks to effectively capture both local patterns and long-term dependencies in code-mixed text. Our model addresses critical challenges in low-resource language processing through a comprehensive preprocessing pipeline and specialized handling of class imbalance and out-of-vocabulary words. Evaluated on a substantial dataset of social media comments, our approach achieved competitive macro F1 scores of 0.3357 for Tamil (**ranked 18**) and 0.3628 for Tulu (**ranked 13**).

Keywords: Sentiment Analysis, Code mixed text, Dravidian languages, Deep Learning, CNN BiLSTM, Tamil-English, Tulu-English.

1 Introduction

It is difficult to analyze sentiment in code-mixed text because of the intricacy of code-switching at many linguistic levels and the absence of annotated datasets. This is particularly relevant for Dravidian languages like Tamil and Tulu, where social media communications frequently mix local languages with English, often written in non native scripts. While recent advances in natural language processing have shown promising results for monolingual text, these systems typically perform poorly on code mixed content, highlighting the need for specialized approaches. Our research addresses these challenges through three main contributions: (1) a novel hybrid deep learning architecture combining CNN BiLSTM networks, specifically designed for code mixed text processing, (2) an effective preprocessing pipeline that handles the unique characteristics of Dravidian code mixed text, and (3) a systematic approach to addressing class imbalance through weighted learning, validated on a diverse dataset comprising over 20,000 Tamil English and 13,000 Tulu English social media comments. **The code is available in this Github Code link.**

2 Related Works

A study from the RANLP 2023 shared task Kanta and Sidorov (2023) et al. examined sentiment study in code mixed dataset. The motive of this study was to categorize YouTube comments into mixed emotions, Neutral, Negative and Positive. The dataset, collected from social media posts, included training, development, and test sets. They used SVM for classification, which got an macro f1 score of 0.147 for tamil-english and 0.518 for tulu-english.

One notable limitation was the low accuracy for tamil-english dataset. For a focused study Hegde et al. (2022) et al. on sentiment examination in data-scarce languages, researchers created a trilingual code mixed Tulu collections with 7,171 YouTube comments. This dataset addresses the lack of tagged data for Tulu. Baseline evaluations using machine learning models showed promising results, though challenges persist due to the informal structure of social media text.

A study by Kannadaguli (2021) et al. focused on Tulu English code mixed text and created the first platinum standard dataset for sentiment analysis. Machine learning and deep learning methods performed better than unsupervised approaches. A recent study by the MUCS team focused on Tamil and Tulu text classification. Using LinearSVC and an ensemble of five classifiers, the team trained models on features derived from word and character n grams.Prathvi et al. (2024).

In a study by Ehsan et al. (2023), sentiment study of code mixed Tulu and Tamil YouTube reviews was tackled using Bidirectional LSTM networks.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 341–345

The models utilized ELMo embeddings fed and trained using larger unannotated code mixed collections for better contextual understanding. Another recent study Tripty et al. (2024). focuses on sentiment study for Tamil and Tulu code mixed text using transformer based models. It found that mBERT and XLM R outperformed others.

This study Chakravarthi et al. (2021). created the multimodal sentiment study dataset for Tamil and Malayalam. They collected YouTube review videos, generated captions, and labeled them for sentiment. The inter annotator consent was verified using Fleiss's Kappa.

Another research by Ponnusamy et al. (2023) tackled sentiment detection in code-mixed social platform comments, which often mix scripts and deviate from grammatical rules. This was achieved using preprocessing and feature extraction techniques along with logistic regression models.

In a recent study, Shetty (2023) et al. overcame the hurdle of sentiment analysis in the tulu dataset. Previous works on code-mixed text demonstrated the effectiveness of machine learning and transformer based models in handling script mixing and linguistic diversity.

However, class imbalance and a lack of annotated datasets remain critical issues for low resource languages. To mitigate these challenges, a new annotated corpus for Tulu was developed and evaluated using standard preprocessing and classification techniques, achieving encouraging results in sentiment classification. This advancement offers a promising foundation that can inspire further work and refinement in low-resource settings.

In another study by Rachana et al. (2023). in which they used fasttext vector representation to train machine learning model for Tulu and Tamil sentences. The models achieved F1-scores of 0.14 for Tamil and 0.204 for Tulu, indicating that there is room for further improvements. These findings underscore the potential benefits of exploring alternative feature representations and tuning strategies to push performance even further.

3 Task Details

Abeera et al. (2023) Social media messages require sentiment analysis because they are mainly code mixed data for dravidian languages. The dataset that we used were code-mixed data Tulu English and Tamil English dataset for the sentiment analysis. Chakravarthi et al. (2020).Lavanya et al. (2024)(Durairaj et al., 2025). The class imbalance issue in the presented dataset illustrates issues that arise in the actual world. The dataset description is displayed in table 1 and 2.

Labels	Train Data	Dev Data	Test Data
NotTulu	4400	543	474
Positive	3769	470	453
Neutral	3175	368	343
Mixed	1114	143	120
Negative	843	118	88

Table 1: Tulu dataset description.

Labels	Train Data	Dev Data	Test Data
Positive	18145	2272	1983
Unknown State	5164	619	593
Mixed Feelings	3662	472	425
Negative	4151	480	458

Table 2: Tamil dataset description.

4 Methodology

Figure 1 represents the proposed workflow, where we consider the raw text data as input. Various preprocessing steps are performed, followed by addressing the class imbalance problem. Later, the target variables will be encoded and later we trained a hybrid model and later we evaluated the model on various evaluation scores.

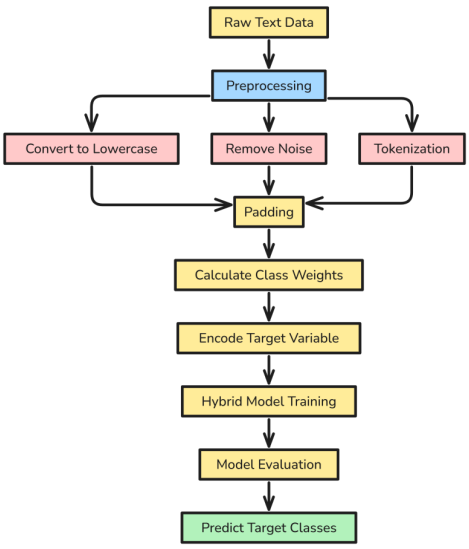

Figure 1: Proposed Methodology.

4.1 Data Preprocessing

In the Initial step, the raw text data was taken, and various preprocessing methods were performed on the text. A function was defined to convert all the text to lowercase. Noise that includes usernames (@), hashtags (#), and punctuation was filtered out from the text using regular expressions. Next, tokenization was performed on the individual words. To handle variable length inputs, each sentence was padded with zeros up to a maximum length to ensure a consistent input size for the model.

The number of unique words in the combined training and validation corpus was used to determine the vocabulary size. To handle words that weren't in the training data, a unique token called <OOV> was utilized.

We used the class weight method in order to address the issue of class inequality. The weight of each class was determined by dividing the total number of samples by the frequency of each class. The label encoding method was ultimately used to transform the target variable into numerical values. The explicit data from the task was utilized for testing and validation, and the dataset from the task was used to train the model.

4.2 Model Architecture

Convolutional and recurrent layers are combined in the design to effectively detect the dataset's long term dependencies as well as local patterns. The model begins with an embedding layer that turns every word into a 100-size dense vector. Word semantic associations can be captured by this layer.

The input sequences are then subjected to a 1D convolutional layer that uses 128 filters with a kernel size of 5 to extract local features. After applying the relu activation function, padding makes sure the output shape is the same as the input shape. To down sample the data, we added a max pooling layer with a pool size of 5.

A bidirectional LSTM layer Kumar et al. (2017). was incorporated to capture long term dependencies. After that, a dense layer using the ReLU activation function with 32 fully connected units comes next. Finally, the output layer, determined by the number of unique tokens, predicts the target classes. The table-3 and 4 tells about the model summary for the tamil and tulu datasets.

4.3 Hyperparameters Setting

The model was configured with the Adam optimizer set to a learning rate of 0.01, and categorical crossentropy was employed as the loss function, making it well-suited for this multi-class classification task. Training was conducted over 100 epochs with a batch size of 32.

Layer	Output Shape	Param
embedding	(None, 40, 100)	2,713,300
conv1d	(None, 40, 128)	64,128
maxpooling1d	(None, 8, 128)	0
bidirectional	(None, 64)	41,216
dense	(None, 32)	2,080
dense	(None, 5)	165

Table 3: Model summary for the tamil dataset.

Layer	Output Shape	Param
embedding	(None, 178, 100)	7,115,400
conv1d	(None, 178, 128)	64,128
maxpooling1d	(None, 36, 128)	0
bidirectional	(None, 64)	41,216
dense	(None, 32)	2,080
dense	(None, 4)	132

Table 4: Model summary for the tulu dataset.

5 Results

5.1 Performance Analysis

The proposed deep learning model was evaluation on differnet metrics which include accuracy, macro f1 score, macro precision score, and macro recall score.

Evaluation Metrics	Scores
Accuracy	0.5114
Macro precision	0.3624
Macro recall	0.3427
Macro f1	0.3357

Table 5: Evaluation scores of the Tamil dataset.

Tables 5 and 6 present the evaluation scores of the Tamil and Tulu datasets. From these tables, it

is observed that the Tulu dataset achieved a macro f1 Score of 0.3628, while the Tamil dataset accomplished a score of 0.3357. Figures 2 and 3 Depict the confusion matrices for the Tamil and Tulu datasets using the proposed model, showing the number of correct predictions along the diagonal across the classes.

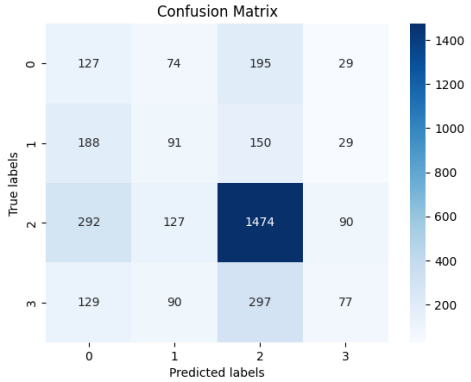

Figure 2: The Tamil dataset's confusion matrix.

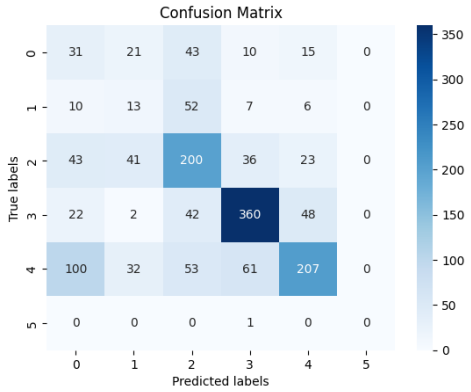

Figure 3: The Tulu dataset's confusion matrix.

Evaluation Metrics	Scores
Accuracy	0.5483
Macro precision	0.3721
Macro recall	0.3676
Macro f1	0.3628

Table 6: Evaluation scores of the Tulu dataset.

6 Conclusion

In this research paper we used tamil-english and tulu-english datasets in which using a combined convolutional and recurrent architecture so that it could capture both local and long term dependencies present in the text. From the results it was observed that for tamil dataset we achieved a macro f1 score of 0.3357 and 0.3628 for tulu dataset.

The OOV problem was addressed using a special <OOV> token, and class imbalance was mitigated through class weighting. Future research will explore the use of pre-trained multilingual models like mBERT and IndicBERT, hyperparameter tuning, and the addition of attention mechanisms to enhance performance. Additionally, comparisons with CNN and LSTM models will be conducted.

7 Limitations

The primary drawback of the suggested approach is the lack of an attention mechanism, which would have enabled the model to concentrate on the crucial segments of the input sequence. Additionally, models like IndicBERT and ModernBERT, which are trained on Indian-context data, could have provided better contextual understanding.

References

S. Kanta and G. Sidorov. 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*, pages 176–179, September 2023.

A. Hegde, M. D. Anusha, S. Coelho, H. L. Shashirekha, and B. R. Chakravarthi. 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, June 2022.

P. Kannadaguli. A code diverse tulu english dataset for nlp based sentiment analysis applications. In *2021 Advanced Communication Technologies and Signal Processing (ACTS)*, pages 1–6. IEEE, December 2021.

B. Prathvi, K. Manavi, K. Subrahmanyapoojary, A. Hegde, G. Kavya, and H. Shashirekha. 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, March 2024.

T. Ehsan, A. Tehseen, K. Sarveswaran, and A. Ali. Alphabrains@ dravidianlangtech: Sentiment analysis of code mixed tamil and tulu by training contextualized elmo word representations. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 152–159, September 2023.

Z. Tripty, M. Nafis, A. Chowdhury, J. Hossain, S. Ahsan, A. Das, and M. M. Hoque. Cuetsentimentsillies@

dravidianlangtech eacl2024: 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*, pages 234–239, March 2024.

B. R. Chakravarthi, K. P. Soman, R. Ponnusamy, P. K. Kumaresan, K. P. Thamburaj, and J. P. McCrae. Dravidianmultimodality: A dataset for multi modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*, 2021.

K. K. Ponnusamy, C. Rajkumar, P. K. Kumaresan, E. Sherly, and R. Priyadharshini. 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, September 2023.

P. Shetty. Poorvi@ dravidianlangtech: Sentiment analysis on code mixed tulu and tamil corpus. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 124–132, September 2023.

K. Rachana, M. Prajnashree, A. Hegde, and H. L. Shashirekha. Mucs@ dravidianlangtech2023: Sentiment analysis in code mixed tamil and tulu texts using fasttext. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 258–265, September 2023.

V. P. Abeera, S. Kumar, and K. P. Soman. Social media data analysis for malayalam youtube comments: Sentiment analysis and emotion detection using ml and dl models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51, September 2023.

B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, and J. P. McCrae. 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, May 2020. Online available: https://aclanthology.org/2020.sltu-1.28.

S. K. Lavanya, A. Hegde, B. R. Chakravarthi, H. L. Shashirekha, R. Natarajan, S. Thavareesan, R. Sakuntharaj, T. Durairaj, P. K. Kumaresan, and C. Rajkumar. 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, March 2024.

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. 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, 2025.

S. S. Kumar, M. A. Kumar, and K. P. Soman. Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets. In *Mining Intelligence and Knowledge Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings 5*, pages 320–334. Springer International Publishing, 2017.