Semantic Analysis on Codemix Text Languages

Submitted by : Swapnil Raj
Roll No: 2024PGCSDS01
NIT,Jamshedpur
2024pgcsds01@nitjsr.ac.in

Under the Guidance of: Dr. Jitesh Pradhan
Associate Professor, Department of CSE, NIT Jamshedpur
Jamshedpur, India

ABSTRACT

The prevalence of code-mixed text on social media, particularly in multilingual regions, poses a significant challenge for traditional sentiment analysis. This project addresses this challenge by building an end-to-end semantic analytics pipeline for Tamil-English code-mixed text. I implemented and compared a wide array of models, starting from classical machine learning baselines like TF-IDF with SVM and Logistic Regression, advancing to deep learning models such as BiLSTM with its variants (CNN and attention mechanisms), and finally evaluating several state-of-the-art multilingual transformer architectures including mBERT, IndicBERT, MuRIL, and XLM-RoBERTa. The results demonstrate a clear performance hierarchy, confirming that multilingual transformers, specifically MuRIL and XLM-RoBERTa, significantly outperform other approaches in this task. This comparative analysis provides a strong benchmark for sentiment analysis on code-mixed Dravidian languages.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; *Neural networks.*

KEYWORDS

Sentiment Analysis, Code-Mixing, Natural Language Processing, Transformer Models, Tamil-English

1 INTRODUCTION

The rise of social media has generated vast amounts of user-generated text, making sentiment analysis a crucial tool for understanding public opinion. In multilingual societies, users often switch between languages, creating code-mixed text that poses a significant challenge for standard NLP models. This project was motivated by the need for effective semantic analysis tools for such languages, specifically focusing on Tamil-English code-mixed data.

The primary objective was to build an end-to-end pipeline for sentiment analysis on this data and to conduct a comparative study of various modeling techniques, from classical machine learning to advanced transformer architectures, to establish which methods are most effective.

2 RELATED WORK

Sentiment analysis has evolved from classical machine learning models using TF-IDF features to more sophisticated deep learning approaches like LSTMs and CNNs. The introduction of transformer models such as BERT and its multilingual variants has further advanced the state-of-the-art. For Indian languages, specialized models like MuRIL [2] and IndicBERT have been developed to better handle the linguistic nuances of the region, including code-mixing. This project builds upon these works by systematically comparing their performance on a specific Tamil-English code-mixed dataset.

3 MATERIALS AND METHODS

3.1 Dataset Description

The study utilized a publicly available dataset of YouTube comments in Tamil-English, annotated with five sentiment labels: *Positive*, *Negative*, *unknown_state*, *Mixed_feelings*, and *not-Tamil*. Due to class imbalance, where some classes had very few samples, categories with fewer than 10 samples were removed to enable robust stratified splitting, resulting in a cleaner dataset for training and evaluation.

3.2 Methodology

I implemented a series of models to compare their predictive performance on the sentiment analysis task. These models can be categorized as follows:

- **Classical Models:** These served as baselines and included TF-IDF vectorization followed by Linear SVM and Logistic Regression classifiers.
- **Deep Learning Models:** This category included several variations of Bidirectional LSTMs (BiLSTM), such as a standard BiLSTM, stacked and dropout variants, a hybrid CNN + BiLSTM, and an Attention-BiLSTM model.
- **Transformer Models:** I evaluated four pre-trained multilingual transformer models: MuRIL, mBERT, IndicBERT, and XLM-RoBERTa. These models were fine-tuned on the task-specific training data.

3.3 Performance Evaluation

The primary metrics for comparing model predictions were the weighted average F1-score, which accounts for class imbalance, and the macro average F1-score, which evaluates performance across all classes equally.

4 RESULTS AND ANALYSIS

The experiments revealed a clear hierarchy in the predictive capabilities of the different model architectures. The overall comparison is detailed below and summarized in Table 1 and Figure 1.

4.1 Comparison of Model Predictions

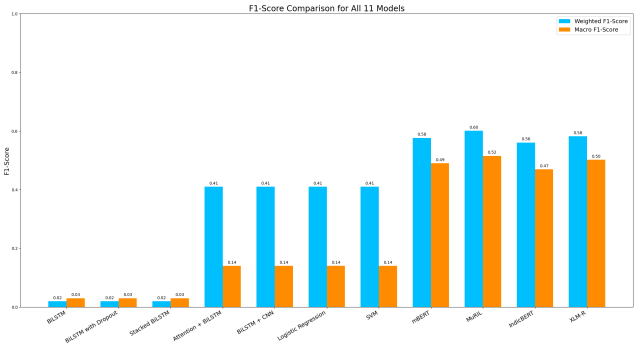
- **Classical Models:** The TF-IDF based models, Logistic Regression and SVM, established the lowest performance benchmarks. Their weighted average F1-scores were both 0.41, limited by their inability to capture deeper contextual nuances.
- **Deep Learning Models:** The BiLSTM models showed varied performance. The more complex hybrid models (CNN + BiLSTM and Attention + BiLSTM) performed identically to

the classical models with weighted F1-scores of 0.41. Interestingly, the simpler BiLSTM variants (Basic, with Dropout, and Stacked) performed poorly on this dataset, achieving a weighted F1-score of only 0.02, indicating they struggled to generalize from the imbalanced data.

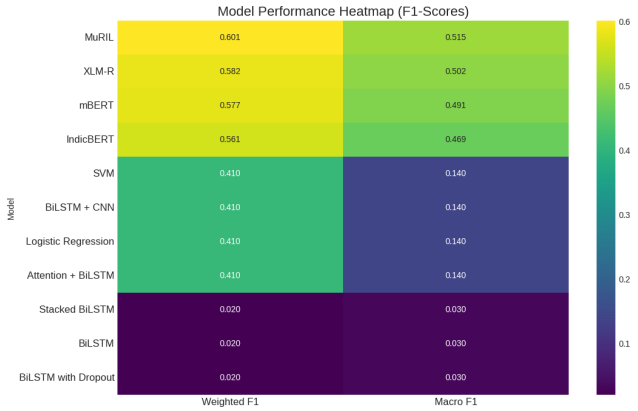
- **Transformer Models:** The transformer-based models significantly outperformed all other approaches, demonstrating their powerful ability to understand complex language patterns. Their predictions were the most accurate.
  - **IndicBERT** and **mBERT** delivered strong results, with weighted average F1-scores of 0.56 and 0.58, respectively.
  - **XLM-RoBERTa** also performed well with a weighted F1-score of 0.58.
  - **MuRIL** emerged as the top-performing model in my study, yielding the highest weighted average F1-score of 0.60 and the highest macro average F1-score of 0.52. This highlights the advantage of its pre-training on a wide array of Indian languages for this code-mixed task.

Table 1: Comparison of model performance on the test set.

Model	Weighted Avg F1	Macro Avg F1
Classical Models		
TF-IDF + SVM	0.41	0.14
TF-IDF + Logistic Regression	0.41	0.14
Deep Learning Models		
Basic BiLSTM	0.02	0.03
BiLSTM with Dropout	0.02	0.03
Stacked BiLSTM	0.02	0.03
CNN + BiLSTM	0.41	0.14
Attention-BiLSTM	0.41	0.14
Transformer Models		
IndicBERT	0.56	0.47
mBERT	0.58	0.49
XLM-RoBERTa	0.58	0.50
MuRIL	**0.60**	**0.52**



(a) Bar graph comparing Weighted and Macro Average F1-Scores for all 11 models.



(b) Heatmap summarizing the performance of all models across key metrics.

Figure 1: Visual comparison of model performance.

4.2 Error Analysis

An analysis revealed that the poor performance of several models was due to extreme class imbalance. The models consistently predicted the majority class ('Positive') and failed to learn the features of minority classes, resulting in low macro average F1-scores. While transformers handled this better, they still showed some confusion between closely related classes like 'Positive' and 'unknown_state'.

5 CONCLUSION

In this project, I successfully built and evaluated an end-to-end semantic analytics pipeline for Tamil-English code-mixed sentiment analysis. My comparative results confirm that multilingual transformer models, particularly those pre-trained on Indian languages like **MuRIL**, significantly outperform other approaches. The extreme class imbalance in the dataset proved to be a major challenge for simpler deep learning models. Future work could focus on techniques to mitigate this imbalance, such as data augmentation or using more advanced loss functions, to further improve model performance.

REFERENCES

[1] Chakravarthi, Bharathi Raja, et al. *Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text.* FIRE 2021.
[2] Khanuja, Simran, et al. *MuRIL: Multilingual Representations for Indian Languages.* arXiv preprint arXiv:2103.1073(2021).