
National Institute of Technology, Jamshedpur

Department of Computer Science Engineering

---

# Semantic Analytics on Code-Mixed Text

*M.Tech Major Project*

---

**Submitted By**

Swapnil Raj

Roll No: 2024PGCSDS01

**Under the Guidance of**

Dr. Jitesh Pradhan

Assistant Professor

Department of CSE, NIT Jamshedpur

September 4, 2025

# Contents

# Declaration

I hereby declare that the work entitled **"Semantic Analytics on Code-Mixed Text"** is my original research work carried out under the supervision of Dr. .Jitesh Pradhan, NIT Jamshedpur This report has not been submitted elsewhere for the award of any degree or diploma.

(Swapnil Raj)

# Acknowledgment

I would like to express my sincere gratitude to my supervisor, Dr. Jitesh Pradhan, for his continuous support, guidance, and insightful feedback throughout this project. I also thank my peers and the Dravidian CodeMix research community for providing valuable datasets and baselines. Finally, I am indebted to my family and friends for their encouragement and motivation.

# Abstract

This project investigates **sentiment analysis on code-mixed Tamil-English text**, a challenging task due to linguistic diversity, transliteration, and noisy social media data.

We benchmarked multiple approaches:

1. **Classical Models:** TF-IDF with Linear SVM and Logistic Regression.
2. **Deep Models:** BiLSTM, CNN+BiLSTM.
3. **Attention-based Hybrids:** Attention-BiLSTM and Attention-BiLSTM+CNN.
4. **Transformer Models:** Google MuRIL, mBERT, IndicBERT, and XLM-RoBERTa.

The dataset used was the FIRE 2021 Dravidian-CodeMix Tamil-English sentiment dataset. The primary evaluation metric was **Weighted F1-score**, chosen to handle class imbalance. Results show that transformer-based approaches significantly outperform classical and deep learning models, with MuRIL and XLM-RoBERTa achieving the highest Weighted F1-scores.

# 1 Introduction

## 1.1 Motivation

In multilingual societies like India, social media users frequently mix languages (code-mixing) within a single sentence. This introduces challenges for natural language processing models. Sentiment analysis on code-mixed data is crucial for opinion mining, hate-speech detection, and customer feedback analysis.

## 1.2 Objectives

- Build a standardized pipeline for sentiment analysis on Tamil-English code-mixed text.
- Compare the performance of classical, deep, hybrid, and transformer-based models.
- Report results using **Weighted F1-score**.
- Identify error patterns unique to code-mixed sentiment analysis.

# 2 Related Work

Previous research in sentiment analysis focused on monolingual English and Hindi corpora. With the introduction of the Dravidian-CodeMix shared tasks at FIRE 2020 and 2021, sentiment analysis of code-mixed Tamil, Malayalam, and Kannada gained momentum. Studies show that while classical models like SVM perform decently, transformer models such as XLM-RoBERTa and MuRIL set new benchmarks.

# 3 Materials and Methods

## 3.1 Dataset Description

We used the FIRE 2021 Dravidian-CodeMix Tamil-English YouTube comments dataset. It contains:

- **Total samples:** 44,020 comments.

- **Classes:** Positive, Negative, Neutral, Mixed feelings, Not-in-intended-language.
- **Splits:** 90% training, 5% validation, 5% testing.

### 3.2 Methodology

**Pipeline:**

1. Data preprocessing: cleaning, normalization, preserving emojis and sentiment cues.
2. Feature extraction: TF-IDF for classical models; embeddings for neural models.
3. Model training: classical ML, BiLSTM/CNN hybrids, attention-based models, transformer fine-tuning.
4. Evaluation: dev/test weighted F1, classification reports, confusion matrices.

### 3.3 Performance Evaluation

The main metric used is the **Weighted F1-score**, defined as:

$$F1_{weighted} = \sum_{i=1}^{L} \frac{n_i}{N} \cdot F1_i$$

where $n_i$ is the number of samples in class $i$, and $N$ is the total number of samples.

## 4 Results

### 4.1 Overall Comparison

### 4.2 Error Analysis

Frequent errors include confusion between *Positive* and *Neutral*, and misclassification of code-switched Romanized Tamil words. Emojis and slang often mislead classical models, but transformers handle them better.

Table 1: Weighted F1-scores of models (dev/test).

| Model | Dev F1 | Test F1 |
|---|---|---|
| TF-IDF + Linear SVM | 0.62 | 0.61 |
| TF-IDF + Logistic Regression | 0.60 | 0.59 |
| BiLSTM | 0.65 | 0.63 |
| CNN + BiLSTM | 0.66 | 0.64 |
| Attention-BiLSTM | 0.67 | 0.65 |
| Attention-BiLSTM + CNN | 0.68 | 0.66 |
| MuRIL | 0.72 | 0.71 |
| mBERT | 0.70 | 0.69 |
| IndicBERT | 0.69 | 0.68 |
| XLM-RoBERTa | 0.74 | 0.73 |

# 5 Conclusion

In this project, we built an end-to-end semantic analytics pipeline for Tamil-English code-mixed sentiment analysis. Our results confirm that multilingual transformers like MuRIL and XLM-RoBERTa outperform other approaches. Classical models remain viable for quick baselines, while attention-based BiLSTMs provide a strong middle ground. Future work includes training code-mix aware transformers, applying data augmentation, and extending experiments to other Dravidian languages.

# Bibliography

# 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:2103.10730, 2021.

[3] Conneau, Alexis, et al. “Unsupervised Cross-lingual Representation Learning at Scale.” ACL 2020.