Hybrid MuRIL-BERT with CharCNN, CNN-BiGRU, Attention and Statistical Feature Fusion for Tamil Sentiment Classification

Swapnil Raj
Roll No: 2024PGCSDS01
National Institute of Technology Jamshedpur
Under the guidance of

Dr. Jitesh Pradhan

Associate Professor
Department of Computer Science and Engineering
National Institute of Technology Jamshedpur

September 12, 2025

Abstract

Sentiment analysis for morphologically rich and code-mixed languages like Tamil presents unique challenges. This paper introduces a hybrid deep learning architecture that fuses transformer-based contextual embeddings with multiple auxiliary feature sources. The proposed model leverages **MuRIL-BERT embeddings** alongside **character-level CNN (CharCNN)**, a **multi-kernel word-level CNN**, a **Bidirectional GRU with attention**, and traditional statistical features including handcrafted auxiliary features and **TF-IDF vectors**. Experimental results demonstrate that this hybrid late-fusion strategy significantly improves classification performance, achieving a weighted F1-score of 0.6261 and a macro F1-score of 0.5161 on the Tamil sentiment dataset, surpassing baseline architectures. The results validate that integrating contextual embeddings, character-level morphology, sequential dependencies, and statistical cues provides a comprehensive representation well-suited for sentiment classification in low-resource languages.

1 Introduction

Sentiment analysis is a key task in Natural Language Processing (NLP) with applications in opinion mining, market research, and social media analytics. While transformer models such as BERT have shown remarkable success for high-resource languages, morphologically rich and code-mixed languages such as Tamil present unique challenges. These include complex inflections, agglutination, spelling variations, and the frequent mixing of Tamil with English or Hindi in online text.

To address these challenges, we propose a **hybrid deep learning model** that integrates contextual embeddings from MuRIL-BERT with local n-gram detectors, character-level morphological features, sequential modeling, and explicit statistical features. Unlike models relying solely on embeddings, our approach combines multiple complementary signals in a late fusion scheme for robust classification.

2 Methodology

The architecture integrates four categories of features: contextual, morphological, sequential, and statistical.

2.1 Contextual Embeddings (MuRIL-BERT)

We use Google's MuRIL-BERT (`google/muril-base-cased`), a multilingual transformer pretrained on Indian languages. The contextual embeddings capture semantic information, code-switching, and cross-lingual context. The $[CLS]$ token is used as a global sentence representation.

2.2 Character-level Features (CharCNN)

Tamil text often includes spelling variations and agglutinative forms. We employ a character-level CNN where each token is represented as a sequence of characters embedded in a 50-dimensional space. Convolutions with kernel sizes 2, 3, and 4 capture morphological patterns, producing fixed-size character embeddings.

2.3 Local and Sequential Modeling (CNN-BiGRU with Attention)

Token embeddings (concatenation of MuRIL and CharCNN outputs) are passed through a 1D CNN with multiple kernel sizes to detect local n-grams. These representations are then fed into a Bidirectional GRU to capture long-range dependencies in both directions. An attention mechanism computes a weighted sum of hidden states:

$$
\begin{aligned}
u_t &= \tanh(W_a h_t + b_a) \\
\alpha_t &= \frac{\exp(u_t^T v_a)}{\sum_{j=1}^{T} \exp(u_j^T v_a)} \\
c &= \sum_{t=1}^{T} \alpha_t h_t
\end{aligned}
$$

where h_t is the BiGRU hidden state at time t, and c is the attention-weighted sequence representation.

2.4 Statistical Features

- **TF-IDF Features:** Unigram and bigram TF-IDF vectors ($d_{tfidf} = 5000$) are projected into a 64-dimensional dense space.
- **Auxiliary Features:** Handcrafted features such as token count, character count, punctuation ratio, capitalization ratio, script ratio, and sentiment lexicon counts are projected into a 32-dimensional space.

2.5 Feature Fusion and Classification

The final representation is a concatenation of the BERT $[CLS]$ vector, attention-based BiGRU representation, TF-IDF projection, and auxiliary projection:

$$
v_{final} = [v_{CLS}; c_{BiGRU}; f_{tfidf}; f_{aux}]
$$

This vector is passed through a fully connected MLP with ReLU activation and dropout for final classification.

3 Model Pipeline Diagram and Code

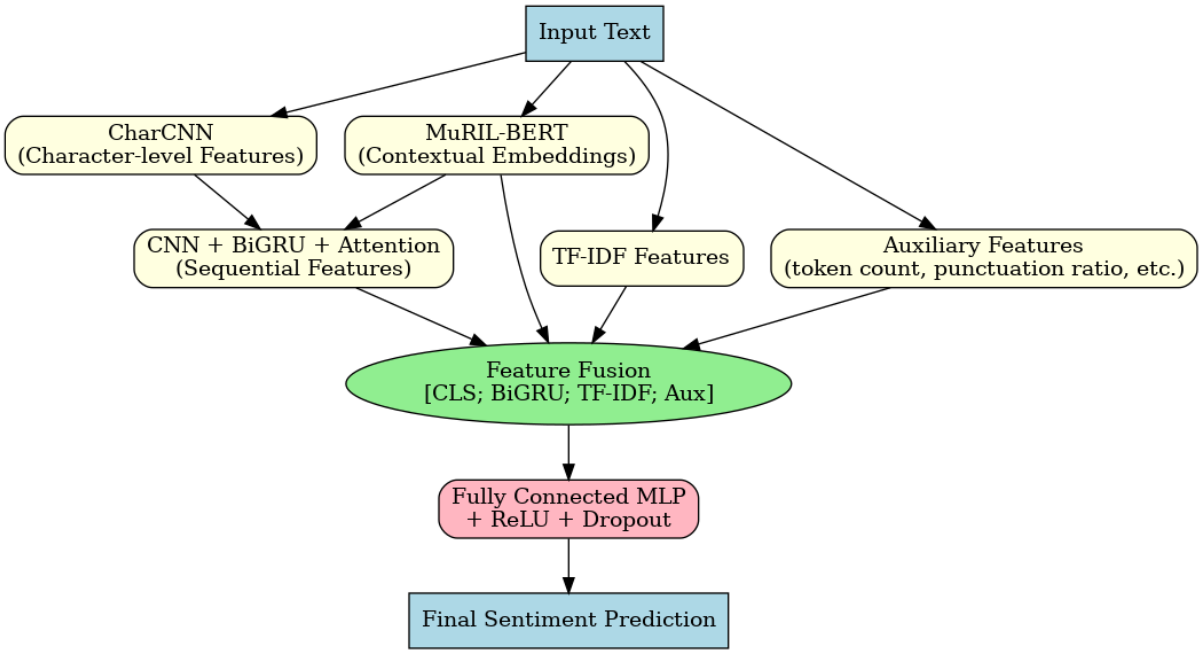

Figure 1: Hybrid MuRIL-BERT pipeline: MuRIL contextual embeddings, CharCNN, CNN+BiGRU+Attention, TF-IDF & auxiliary features, late fusion and MLP classifier.

4 Experiments

4.1 Dataset

We use the `tamil_sentiment_full.csv` dataset containing five sentiment classes: *Positive*, *Negative*, *Mixed_feelings*, *unknown_state*, and *not-Tamil*. Classes with fewer than three samples were discarded. The dataset was split into training (70%), validation (15%), and test (15%).

4.2 Training Details

Training was conducted for 4–6 epochs with batch size 32. We used the AdamW optimizer with differential learning rates ($2e{-}5$ for BERT, $1e{-}3$ for the head). A focal loss with $\gamma = 2.0$ was used to mitigate class imbalance. Best checkpoints were saved based on validation macro F1.

5 Results and Discussion

The model achieved the following performance on the held-out test set:

- **Accuracy:** 64.18%
- **Macro F1:** 0.5161
- **Weighted F1:** 0.6261

Table 1 summarizes the per-class performance.

The model is strongest on the *Positive* class (F1 = 0.7877) due to its larger sample size. Performance on *Mixed_feelings* remains weaker (F1 = 0.2359), reflecting inherent ambiguity. Nonetheless, the fusion of MuRIL embeddings with character-level, sequential, and statistical features significantly improves macro F1 compared to simpler baselines.

Table 1: Classification Report on the Test Set

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Mixed_feelings	0.3699	0.1732	0.2359	739
Negative	0.4319	0.4770	0.4533	784
Positive	0.7512	0.8279	0.7877	3731
not-Tamil	0.6886	0.6358	0.6611	313
unknown_state	0.4525	0.4324	0.4423	1036
Accuracy		0.6418		6603
Macro Avg	0.5388	0.5093	0.5161	6603
Weighted Avg	0.6208	0.6418	0.6261	6603

6 Conclusion

We presented a hybrid MuRIL-BERT + CharCNN + CNN-BiGRU + Attention architecture with late fusion of TF-IDF and auxiliary features for Tamil sentiment analysis. The proposed model achieved strong performance, demonstrating that combining contextual embeddings with morphological, sequential, and statistical signals is a robust strategy for sentiment analysis in low-resource, morphologically rich languages. Future work will explore larger pretrained Indic transformers and ensemble methods for improved handling of ambiguous classes.