# byteSizedLLM@DravidianLangTech 2025: Sentiment Analysis in Tamil Using Transliteration-Aware XLM-RoBERTa and Attention-BiLSTM

**Durga Prasad Manukonda**
ASRlytics
Hyderabad, India
mdp0999@gmail.com

**Rohith Gowtham Kodali**
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

## Abstract

This study investigates sentiment analysis in code-mixed Tamil-English text using an Attention BiLSTM-XLM-RoBERTa model, combining multilingual embeddings with sequential context modeling to enhance classification performance. The model was fine-tuned using masked language modeling and trained with an attention-based BiLSTM classifier to capture sentiment patterns in transliterated and informal text. Despite computational constraints limiting pretraining, the approach achieved a Macro f1 of 0.5036 and ranked first in the competition. The model performed best on the Positive class, while Mixed Feelings and Unknown State showed lower recall due to class imbalance and ambiguity. Error analysis reveals challenges in handling non-standard transliterations, sentiment shifts, and informal language variations in social media text. These findings demonstrate the effectiveness of transformer-based multilingual embeddings and sequential modeling for sentiment classification in code-mixed text.

## 1 Introduction

Sentiment analysis involves identifying subjective opinions or emotions in text and has gained significant attention in both academia and industry. With the rise of social media, sentiment detection in Dravidian languages has become increasingly relevant, especially given the prevalence of code-mixing. Code-mixed texts, often written in non-native scripts, pose challenges for traditional monolingual sentiment analysis models due to complex linguistic variations and switching between languages.

The Shared Task on Sentiment Analysis in Tamil and Tulu at DravidianLangTech@NAACL 2025 focuses on message-level polarity classification of code-mixed Tamil-English and Tulu-English texts. Given a YouTube comment or post, the goal is to classify it as positive, negative, neutral, or mixed sentiment. The dataset, collected from social media, presents real-world challenges such as class imbalance and linguistic variability, necessitating robust NLP techniques for effective classification.

To address these challenges, we propose a transliteration-aware fine-tuning approach using XLM-RoBERTa, a state-of-the-art multilingual transformer model. The model is fine-tuned using Masked Language Modeling (MLM) on a subset of the AI4Bharath dataset(Kunchukuttan et al., 2020), incorporating original, fully transliterated, and partially transliterated text. This pretraining strategy equips the model to handle native scripts, Romanized text, and mixed-script data effectively.

Additionally, we integrate XLM-RoBERTa embeddings into a hybrid architecture with a attention-BiLSTM. The embeddings are projected and refined to capture complex contextual relationships in multilingual text. Dropout regularization and gradient clipping ensure stable training. Our approach achieves state-of-the-art performance, demonstrating the effectiveness of transliteration-aware pretraining and hybrid architectures in handling sentiment classification for code-mixed Dravidian languages.

This study analyzes data preprocessing, MLM training, and classifier design, introducing innovations that improve detection accuracy and scalability. The proposed framework enhances Sentiment Analysis in Tamil, providing insights into model performance and deployment challenges.

## 2 Related Work

Sentiment Analysis on social media has progressed significantly, with growing attention to low-resource languages like Tamil. Chakravarthi et al. (2021) and B et al. (2022) organized shared tasks to promote Sentiment Analysis in code-mixed Dravidian languages, laying a foundation for tackling linguistic diversity in Sentiment Analysis.



*Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 212–217

Several approaches have explored Tamil-English code-mixed data. S R et al. (2022) addressed data imbalance using kernel-based learning and advanced feature selection techniques, while Shanmugavadivel et al. (2022) employed hybrid deep learning models, combining CNN and BiLSTM architectures, achieving strong results for mixed-language datasets. Preprocessing steps, including emoji and punctuation removal, and TF-IDF-based feature extraction, were crucial to their success.

Sentiment Analysis in Tamil has been explored through various shared tasks, such as DravidianLangTech@RANLP 2023 (Priyadharshini et al., 2023; Hegde et al., 2023a) and DravidianLangTech@EACL 2024 (Sambath Kumar et al., 2024). In 2023, XLM-RoBERTa with adversarial and ensemble training demonstrated the effectiveness of transformers for Tamil-English code-mixed text (Luo and Wang, 2023). The task also addressed abusive language detection in Tamil, Telugu, and Tamil-English code-mixed texts using approaches like LinearSVC with n-grams and Transfer Learning models with BERT variants, emphasizing the ongoing challenges in handling abusive content effectively (Hegde et al., 2023b).

In 2024, B et al. (2024) implemented SVM and an ensemble of ML classifiers—Support Vector Model (SVM), Random Forest (RF), and k Nearest Neighbors (kNN)—for Tamil-English code-mixed sentiment analysis. They used GridSearch for hyperparameter tuning, achieving a top macro F1 score and securing a top rank in the shared task.

Despite these advancements, Sentiment Analysis in Tamil remains an open challenge, requiring further improvements in methods and performance.

## 3 Dataset

The dataset for Sentiment Analysis in Tamil task consists of code-mixed Tamil-English comments and posts collected from social media platforms. Each instance is annotated with one of four sentiment labels: Positive(0), Negative(1), Mixed Feelings(2), and Unknown State(3). The dataset presents class imbalance, reflecting real-world sentiment distribution in online discourse(Chakravarthi et al., 2020).

The data is divided into training, validation, and test sets, ensuring a robust benchmark for sentiment classification. This is a message-level polarity classification task, and Table 1 summarizes the dataset distribution.

| Label | Train | Val | Test | Total |
|---|---|---|---|---|
| 0 | 18145 | 2272 | 1983 | 22400 |
| 1 | 4151 | 480 | 458 | 5089 |
| 2 | 3662 | 472 | 425 | 4559 |
| 3 | 5164 | 619 | 593 | 6376 |
| **Total** | 31122 | 3843 | 3459 | 38424 |

Table 1: Dataset distribution across sentiment labels in Train, Validation, and Test splits.

This dataset serves as a benchmark for exploring sentiment expression in code-mixed Tamil-English text, addressing challenges such as transliteration, informal language, and code-switching in social media discourse.

## 4 Models

This section presents the models used in our experiments. Fine-tuned XLM-RoBERTa with Masked Language Modeling (MLM) enhances processing of Tamil-English code-mixed text. An attention-driven BiLSTM further refines embeddings, improving contextual understanding and sequence modeling.

### 4.1 Fine-Tuning XLM-RoBERTa with MLM

XLM-RoBERTa, a multilingual transformer model based on RoBERTa, is trained on a large-scale Common Crawl corpus spanning 94 languages (Conneau et al., 2019). It employs dynamic masking and optimized pretraining, enabling it to capture complex linguistic patterns across languages.

To enhance its ability to process transliterated and code-switched Tamil-English text, we fine-tuned the base XLM-RoBERTa model using Masked Language Modeling (MLM). This pretraining strategy involves masking random tokens and training the model to predict them, allowing it to learn robust contextual embeddings tailored to bilingual text.

The MLM training dataset was constructed from monolingual Tamil social media text, fully transliterated text in Roman script, and partially transliterated text with 20–70% of words transliterated. This approach enabled the model to recognize native script, Romanized text, and mixed-script data, crucial for processing real-world Tamil-English social media content.

The fine-tuned XLM-RoBERTa model (TamilXLM_Roberta[1]) serves as the embedding

[1] https://huggingface.co/bytesizedllm/TamilXLM_

backbone for sentiment classification, enhancing its ability to handle linguistic and orthographic variability in code-mixed datasets.

### 4.2 Attention BiLSTM-XLM-RoBERTa Model

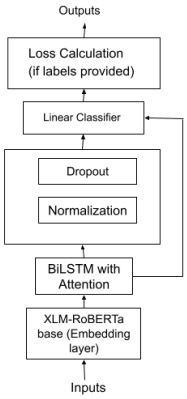


Figure 1: Architecture of the BiLSTM-XLM-RoBERTa Classifier Model.

This study introduces a hybrid Attention BiLSTM-XLM-RoBERTa model for multi-label classification, integrating fine-tuned XLM-RoBERTa embeddings with a BiLSTM and attention mechanism (Liu and Guo, 2019; Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Kodali et al., 2025; Manukonda and Kodali, 2025, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b). As shown in Figure 1, the model captures contextual dependencies using BiLSTM and assigns dynamic importance to hidden states via attention.

XLM-RoBERTa generates contextual embeddings, which are processed by BiLSTM to extract forward and backward hidden states. An attention mechanism computes weight distributions to refine the representation:

$$\mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t, \quad \alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T} \exp(\mathbf{a}_t)} \tag{1}$$

Residual components such as layer normalization and dropout are applied to the attention-weighted representation to stabilize training and reduce overfitting:

Roberta

$$\mathbf{H}_{dropout} = Dropout(LayerNorm(\mathbf{H}_{attended})) \tag{2}$$

Finally, a classification layer outputs logits:

$$\textbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \tag{3}$$

The model is trained using cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{4}$$

This architecture effectively combines XLM-RoBERTa embeddings, BiLSTM, and attention to enhance multi-label classification in code-mixed text.

## 5 Experiment Setup

The experiments evaluate the integration of attention-based BiLSTM with fine-tuned XLM-RoBERTa embeddings for sentiment analysis in code-mixed Tamil-English text. XLM-RoBERTa was fine-tuned using Masked Language Modeling (MLM) with a 15% masking probability, a batch size of 16, and a learning rate of $5 \times 10^{-5}$. The model was trained for up to ten epochs with early stopping based on validation perplexity.

For classification, the fine-tuned embeddings were processed through a BiLSTM model with two LSTM layers (hidden size 512) and an attention mechanism to enhance contextual representation. A dropout probability of 0.3 was applied for generalization. The model was trained using AdamW with a learning rate of $2.5 \times 10^{-5}$ and weight decay of 0.01, running for six epochs with early stopping based on validation loss and macro F1-score.

This setup demonstrates the effectiveness of combining XLM-RoBERTa embeddings, BiLSTM, and attention mechanisms for sentiment classification in Tamil-English text, addressing challenges such as transliteration, informal language, and linguistic variability in social media data.

## 6 Results and Discussion

XLM-RoBERTa achieved a perplexity of 4.9 for Tamil bilingual text, indicating its effectiveness in modeling code-mixed language representations.

The performance of the Attention BiLSTM-XLM-RoBERTa model was evaluated on the code-mixed Tamil-English sentiment analysis task[2]. The

[2] https://github.com/mdp0999/Sentiment-Analysis-in-Tamil

| Label | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Mixed Feelings | 0.27 | 0.24 | 0.26 | 425 |
| Negative | 0.53 | 0.46 | 0.49 | 458 |
| Positive | 0.76 | 0.84 | 0.79 | 1983 |
| Unknown State | 0.51 | 0.41 | 0.45 | 593 |
| **Accuracy** | - | - | 0.64 | 3459 |
| **Macro Avg** | 0.52 | 0.49 | 0.50 | 3459 |
| **Weighted Avg** | 0.62 | 0.64 | 0.63 | 3459 |

Table 2: Classification Report on the Test Set for Sentiment Analysis in Code-Mixed Tamil-English Text

classification report in Table 2 shows an overall accuracy of 64 percent, with a macro F1-score of 0.50 and a weighted F1-score of 0.63.

The model performed best on the Positive class, achieving an F1-score of 0.79. This can be attributed to the higher representation of Positive instances in the dataset, allowing the model to learn its distinguishing features more effectively. In contrast, the Mixed Feelings and Unknown State categories had lower F1-scores of 0.26 and 0.45, respectively, suggesting difficulty in distinguishing ambiguous sentiment. The Negative class obtained a moderate F1-score of 0.49, reflecting challenges in identifying negative sentiment, which often overlaps with neutral or mixed sentiments.

Several misclassifications stemmed from class imbalance, ambiguous sentiment expressions, and code-switching complexity. Mixed Feelings and Unknown State were often misclassified as Positive or Negative due to overlapping linguistic cues, especially in subtle or sarcastic expressions. Errors also arose from transliteration inconsistencies, as Tamil-English text lacks standardized spelling. Variations in transliteration, spelling errors, and informal language led to confusion, affecting sentiment assignment. The model also struggled with sentiment shifts in longer sentences, resulting in incorrect predictions when sentiment changed mid-sentence.

| Team | Score | Rank |
|---|---|---|
| **byteSizedLLM** | **0.5036** | **1** |
| ET2025 | 0.4986 | 2 |
| Hermes | 0.4957 | 3 |
| JustATalentedTeam | 0.4919 | 4 |
| Lemlem | 0.4709 | 5 |

Table 3: Performance ranking of different teams based on their submitted runs.

## 7 Limitations and Future Work

This study was limited by computational constraints, restricting XLM-RoBERTa pretraining and its generalization to diverse Tamil-English code-mixed patterns. Class imbalance, particularly in the Mixed Feelings and Unknown State categories, led to biased classification. Additionally, low transliteration accuracy introduced inconsistencies in text representation, affecting sentiment detection. Addressing these challenges through data augmentation techniques such as back-translation and oversampling could improve recall for underrepresented classes.

Future work will focus on developing more accurate transliteration models for code-mixed text, improving representation consistency. Expanding pretraining on larger datasets and overcoming computational limitations could enhance model performance. Additionally, integrating multimodal data and exploring domain adaptation techniques may improve robustness in handling informal and noisy social media text.

## 8 Conclusion

This study presented an Attention BiLSTM-XLM-RoBERTa model for sentiment analysis in code-mixed Tamil-English text, effectively capturing sentiment cues by leveraging multilingual embeddings and sequential modeling. The model achieved competitive performance, but challenges such as class imbalance, ambiguous sentiment transitions, and low transliteration accuracy affected classification of underrepresented categories. Error analysis highlighted the need for improved handling of informal and transliterated text. Future enhancements, including better pretraining, data augmentation, and robust transliteration models, can further refine sentiment detection in code-mixed social media text.

## References


Prathvi B, Manavi K, Subrahmanyapoojary K, Asha Hegde, Kavya G, and Hosahalli Shashirekha. 2024. 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, St. Julian's, Malta. Association for Computational Linguistics.

Premjith B, Bharathi Raja Chakravarthi, Malliga Subramanian, Bharathi B, Soman Kp, Dhanalakshmi V, Sreelakshmi K, Arunaggiri Pandian, and Prasanna Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 254–260, Dublin, Ireland. 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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2021. Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '20, page 21–24, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, Lavanya S K, Thenmozhi D., Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023a. 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, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Asha Hegde, Kavya G, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2023b. MUCS@DravidianLangTech2023: Leveraging learning models to identify abusive comments in code-mixed Dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 266–274, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Zhipeng Luo and Jiahui Wang. 2023. DeepBlueAI@DravidianLangTech-RANLP 2023. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 171–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and

streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mithun Kumar S R, Lov Kumar, and Aruna Malapati. 2022. Sentiment analysis on code-switched Dravidian languages with kernel based extreme learning machines. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 184–190, Dublin, Ireland. Association for Computational Linguistics.

Lavanya Sambath Kumar, Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, Prasanna Kumar Kumaresan, and Charmathi Rajkumar. 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*, pages 62–70, St. Julian's, Malta. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Comput. Speech Lang.*, 76(C).