Abusive Content Detection in Telugu-English Code-Mixed Social Media Using Hybrid Transformer Architectures

Bojja Revanth Reddy, Sivaiah Bellamkonda

Abstract

The rapid growth of social media has led to a rise in abusive and offensive content, especially in code-mixed languages like Telugu-English. Detecting such content is challenging because of transliteration, spelling variations, informal writing styles, and frequent switching between languages. This paper presents a hybrid deep learning model for abusive content detection in Telugu-English code-mixed comments. A pretrained transformer model, DeBERTa, is used to generate contextual embeddings, which are further processed through parallel Multi-Channel CNN and BiLSTM networks for feature extraction. The CNN captures important local patterns and phrases, while the BiLSTM learns sequential and contextual information from the text. The combined features are then used for final classification. The proposed model effectively handles the linguistic complexity of Telugu-English code-mixed social media data and improves abusive content detection performance.

1 Introduction

Social media platforms have transformed global communication by enabling real-time interaction and large-scale content sharing. While these platforms facilitate engagement and information exchange, they also enable the spread of abusive language, abusive speech, cyberbullying, and misinformation. Due to the massive volume of user-generated content produced daily, manual moderation is infeasible, making automated detection systems essential.

Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques are widely adopted for automated abusive content detection. However, detecting abusive language in Telugu-English code-mixed social media text presents additional challenges. Code-mixing refers to the blending of two languages within a single sentence or discourse. In Telugu-English contexts, users frequently write Telugu words using the Roman script, resulting in inconsistent transliterations and spelling variations.

For example, a single Telugu word may appear in multiple Romanized forms depending on user preference. Informal grammar, slang, repeated characters, and creative spellings further complicate tokenization and feature extraction. Additionally, contextual ambiguity and sarcasm make it difficult to distinguish between genuine and abusive intent.

Traditional NLP systems trained on monolingual corpora struggle to generalize in such multilingual and noisy environments. Therefore, robust preprocessing techniques, effective feature representations, and contextual deep learning models are necessary for reliable abusive content detection in Telugu-English code-mixed data.

2 Literature Survey

Recent research in abusive content detection has evolved from keyword-based filtering to machine learning and deep learning approaches. Traditional supervised models such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression rely on feature engineering techniques like n-grams and TF-IDF representations. These methods provide strong baseline performance but often fail to capture contextual nuances in complex language settings.

Deep learning models, including CNNs and LSTMs, improve performance by learning hierarchical representations of text. More recently, transformer-based architectures such as BERT, XLM-RoBERTa, and IndicBERT have demonstrated superior performance in multilingual and low-resource scenarios. These models leverage subword tokenization and self-attention mechanisms to capture contextual dependencies and handle spelling variations.

However, Telugu-English code-mixed abusive

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 1–5

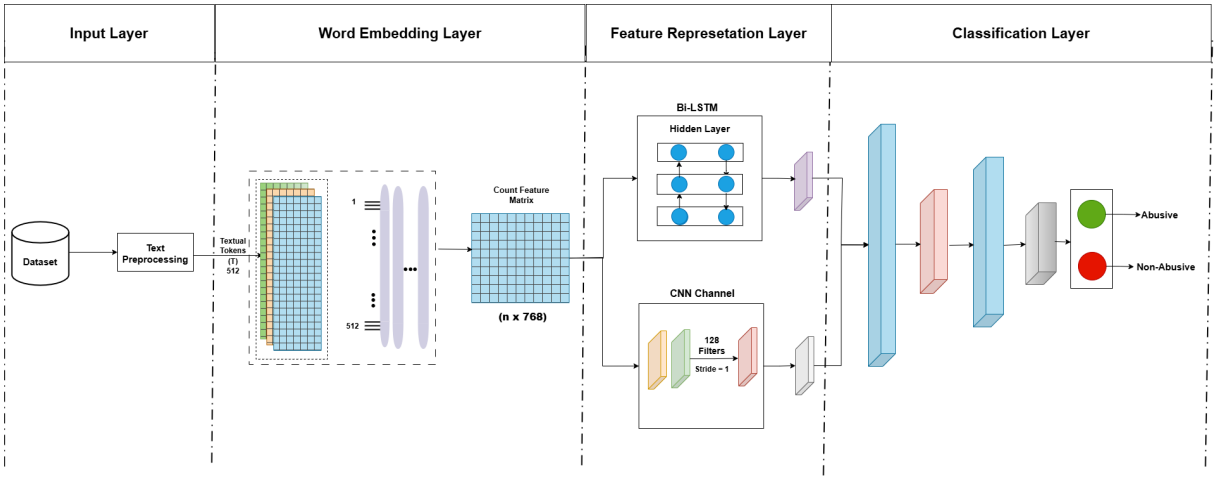

Figure 1: Hybrid transformer–CNN–BiLSTM architecture.

detection remains underexplored due to limited annotated datasets, transliteration variability, and morphological complexity of Telugu. This gap highlights the need for domain-specific preprocessing and fine-tuned multilingual transformer models tailored to code-mixed environments.

3 Methodology

3.1 Preprocessing

Preprocessing was performed to ensure textual consistency and quality before model training. Telugu-English code-mixed social media text often contains noise, informal expressions, spelling variations, and inconsistent transliterations. Therefore, multiple normalization steps were applied. Missing values, URLs, emojis, and extra whitespace were removed during data cleaning. Special symbols and unnecessary punctuation marks were eliminated to reduce noise. Common Telugu-English stopwords were filtered out to minimize redundancy and enhance discriminative learning.

To address lexical inconsistency, variant spellings and abusive term synonyms were standardized to maintain uniform representation across samples. Furthermore, after data augmentation, native Telugu script was converted back into Romanized Telugu script to ensure consistent feature extraction and compatibility with the model tokenizer. These preprocessing steps ensured uniform textual representation and improved interpretability and downstream classification performance.

3.2 Dataset Augmentation

To improve linguistic diversity and model robustness, data augmentation techniques were applied. The final augmented dataset consists of 49,910 comments, providing broader coverage of both abusive and non-abusive expressions. Augmentation included synonym substitution, controlled word insertion and deletion, and generation of spelling and dialectal variants commonly observed in Telugu-English code-mixed social media text. These techniques simulate natural user behavior and enhance the model's ability to generalize to unseen variations of abusive language.

3.3 Model Architecture

The proposed model integrates transformer-based contextual embeddings with convolutional and recurrent neural networks to capture both local and global linguistic patterns. A multilingual transformer encoder first converts input text into contextualized embeddings, capturing semantic relationships and multilingual characteristics of code-mixed data.

The contextual embeddings are then passed through a Convolutional Neural Network (CNN) layer to extract local n-gram and phrase-level patterns using convolution and max-pooling operations. In parallel, a Bidirectional Long Short-Term Memory (BiLSTM) network processes the embeddings to capture long-range sequential dependencies by analyzing text in both forward and backward directions.

The outputs from the CNN and BiLSTM layers are concatenated to form a rich fused representation that combines local lexical features and global contextual information. This fused feature vector is passed through a fully connected dense layer with a Sigmoid activation function to perform binary clas-

sification and predict whether the input comment is abusive or non-abusive.

This hybrid architecture enables effective modeling of contextual semantics, phrase-level patterns, and sequential dependencies, making it well-suited for abusive content detection in Telugu-English code-mixed social media text.

4 Experiment

This section presents the experimental configuration, dataset details, and evaluation metrics used to assess the performance of the proposed abusive content detection model for Telugu-English code-mixed social media text.

4.1 Experimental Setup

All experiments were conducted using Jupyter Notebook, which provides an integrated environment for coding, visualization, and model evaluation. The models were implemented using the TensorFlow framework with Keras as the high-level API for constructing neural network architectures.

The proposed hybrid architecture integrates a multilingual transformer encoder with CNN and BiLSTM layers. Transformer-based models such as XLM-RoBERTa and IndicBERT were used to generate contextual embeddings. The dataset was divided into 75% training data and 25% testing data to ensure reliable evaluation of generalization performance.

The models were trained using binary cross-entropy loss with the Adam optimizer. Early stopping was employed to prevent overfitting and ensure stable convergence.

4.2 Dataset

The dataset consists of Telugu-English code-mixed social media comments annotated for abusive (abusive) and non-abusive (non-abusive) categories. The data contains transliterated Telugu text written in Roman script along with English words, reflecting realistic informal online communication.

After preprocessing and augmentation, the final dataset contains a total of **49,910 comments**. The class distribution of the final augmented dataset is shown in Table 1.

The augmented dataset provides improved linguistic diversity and balanced representation of abusive and non-abusive expressions, enabling better generalization of the classification model.

Table 1: Final Augmented Dataset Distribution

Class	Label	Number of Samples
Non-abusive	0	23,342
Abusive	1	26,568
Total		**49,910**

4.3 Performance Metrics

The models were evaluated using standard classification metrics, including accuracy, precision, recall, and F1 score.

1. **Accuracy**:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

2. **Precision**:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

3. **Recall**:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

4. **F1 Score**:

$$F1 = 2 \cdot \left(\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \right) \tag{4}$$

Here, **TP** denotes True Positives, **TN** denotes True Negatives, **FP** denotes False Positives, and **FN** denotes False Negatives. Since abusive content detection may involve class imbalance, macro-averaged F1 score is emphasized to ensure fair performance evaluation across both classes.

5 Results

This section presents the quantitative evaluation and analytical insights of the proposed hybrid architecture for abusive content detection in Telugu-English code-mixed text.

5.1 Quantitative Performance

To evaluate the impact of different embedding backbones combined with additional feature extraction layers, we compare transformer-based contextual encoders integrated with parallel CNN–BiLSTM modules. The CNN component captures local phrase-level patterns, while the BiLSTM models long-range contextual dependencies.

Table 2 reports the performance metrics on the test dataset.

Table 2: Performance Comparison of Embedding Backbones with Parallel CNN–BiLSTM Feature Extraction

Embedding Layer	Feature Extraction	Accuracy	Precision	Recall	F1-score
mBERT	CNN–BiLSTM	0.86	0.86	0.86	0.86
IndicBERT	CNN–BiLSTM	0.87	0.87	0.86	0.87
DeBERTa	CNN–BiLSTM	**0.97**	**0.97**	**0.97**	**0.97**

Among the evaluated configurations, DeBERTa combined with parallel CNN–BiLSTM layers achieves the highest performance across all metrics, reaching an F1-score of 0.97. mBERT and IndicBERT provide competitive yet comparatively lower performance, indicating that contextual embedding strength plays a crucial role in modeling code-mixed abusive expressions.

5.2 Discussion

The performance gap between DeBERTa and the other embedding backbones highlights the importance of enhanced contextual representation for noisy and transliterated social media text. Telugu-English code-mixed data often contains inconsistent spelling patterns, informal grammar, and implicit abusive intent. Stronger attention mechanisms and refined positional encoding strategies appear to better capture such linguistic variability.

The consistent alignment between precision and recall for DeBERTa suggests balanced classification behavior without overfitting toward either abusive or non-abusive classes. This balance is critical in abusive content detection, where false negatives may allow harmful content to pass undetected, and false positives may incorrectly flag benign content.

Furthermore, the integration of CNN and BiLSTM layers contributes to improved robustness by simultaneously capturing short abusive phrases and broader contextual meaning. The results confirm that hybrid feature extraction combined with advanced transformer embeddings significantly enhances performance in low-resource, code-mixed environments.

6 Conclusion

This paper addressed the problem of abusive content detection in Telugu-English code-mixed social media text. We proposed a hybrid architecture that combines transformer-based contextual embeddings with parallel CNN and BiLSTM layers to effectively capture both local lexical cues and long-range contextual dependencies.

Experimental results demonstrate that the embedding backbone plays a critical role in performance. Among the evaluated models, DeBERTa integrated with CNN–BiLSTM feature extraction achieved the best overall results, significantly outperforming other transformer backbones. The balanced precision and recall values indicate stable and reliable classification behavior across abusive and non-abusive classes.

The findings confirm that integrating strong contextual encoders with complementary feature extraction mechanisms enhances robustness in handling transliterated, noisy, and morphologically complex code-mixed data.

7 Future Work

Although the proposed approach achieves strong performance, several research directions can further improve abusive content detection in code-mixed environments. Future work can focus on domain-adaptive pretraining using large-scale Telugu-English social media corpora to better capture dialectal variations and evolving abusive expressions.

Additionally, advanced fusion strategies such as attention-based feature interaction may improve integration between CNN and BiLSTM representations. Exploring lightweight transformer variants could also support real-time deployment in large-scale moderation systems.

Extending the task to multi-class abusive categorization and applying cross-lingual transfer learning across other Indian code-mixed languages are promising directions. Incorporating explainability techniques will further enhance transparency and trust in automated content moderation systems.

References

[1] Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors

with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[3] Bharathi Raja Chakravarthi and R. Priyadharshini. 2020. A dataset for Tamil-English code-mixed offensive text classification. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, et al. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[5] J. Angel Diaz-Garcia and Joao Paulo Carvalho. 2025. A survey of textual cyber abuse detection using cutting-edge language models and large language models. *arXiv preprint arXiv:2501.05443*.

[6] P. Jakkula. 2020. Comparative study on Telugu text classification using machine learning and deep learning techniques. *SciSpace Link*.

[7] V. Khanduja, S. Singh, and H. Bansal. 2024. Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation. *Systems and Soft Computing*, 6:200112.

[8] A. Kumar, A. Ojha, and T. Solorio. 2021. Code-mixed abusive language detection: Challenges and solutions. Unpublished manuscript.

[9] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*.

[10] S. Maddu and V. R. Sanapala. 2024. A survey on NLP tasks, resources and techniques for low-resource Telugu-English code-mixed text. *ACM Digital Library*.

[11] Thomas Mandl, Sandip Modha, Punyajoy Majumder, Diptesh Patel, et al. 2019. Overview of the HASOC track at FIRE 2019. In *Proceedings of FIRE*, pages 14–17.

[12] Atul Krushna Puranik, Ritesh Kumar, Navneet Kumar, Bharathi Raja Chakravarthi, and Bornini Lahiri. 2021. Telugu-English code-mixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 179–190.

[13] Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, pages 133–142.

[14] Tharindu Ranasinghe, Marcos Zampieri, and Constantin Orasan. 2020. MUDES: Multilingual detection of offensive spans. In *Proceedings of EMNLP*, pages 6222–6231.

[15] A. Singh and S. Rajput. 2021. Code-mixed sentiment classification using transformer models. In *ICON 2021: 18th International Conference on Natural Language Processing*.

[16] Akarsh Srivastava and Ankush Singh. 2020. PHINC: A parallel Hindi-English code-mixed corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1481–1488.

[17] Nafisa Tabassum, Mosabbir Khan, Shawly Ahsan, Jawad Hossain, and Mohammed Moshiul Hoque. 2024. Hate and offensive language detection in Telugu code-mixed text using transliteration-augmentation. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 167–172.

[18] Kusampudi Varma, Preetham Sathineni, and Radhika Mamidi. 2021. Sentiment analysis in code-mixed Telugu-English text with unsupervised data normalization. Pages 753–760.

[19] P. Vijayaraghavan and A. Das. 2022. Abusive speech detection using deep learning models. Unpublished manuscript.