Enhanced BERT-BiLSTM Model with Attention Mechanism for Robust Offensive Language Detection

Pranjali Pachpor, Nitesh Gupta, Anurag Shrivastava

Abstract:

Offensive language detection is a vital challenge in natural language processing (NLP), particularly on online platforms where harmful content spreads rapidly. This paper introduces a robust hybrid deep learning model that combines BERT-based contextual embeddings with Bidirectional LSTM (BiLSTM) and an attention mechanism to capture both semantic depth and sequential dependencies. The proposed framework emphasizes key offensive triggers within text, enabling precise detection. Experimental results demonstrate superior performance, with 95.36% accuracy, 94.87%, outperforming baseline BERT and existing models. The approach generalizes well across diverse datasets, handles imbalanced data efficiently, and provides a scalable solution for automated offensive language moderation, offering a significant step toward safer and more responsible online communication.

Keywords— Offensive language detection, Social media, Multilingual, Transfer learning, Text classification, Natural language processing

I. INTRODUCTION

The rapid expansion of social media platforms has led to an unprecedented increase in user-generated content. While these platforms have enhanced global connectivity, they have also become fertile grounds for the spread of offensive, abusive, and hate-filled language. Detecting such content automatically is a significant challenge in Natural Language Processing (NLP) due to the complexity of linguistic nuances, slang, sarcasm, and multilingual communication styles [1], [2]. Effective offensive language detection systems are therefore crucial for maintaining healthy online interactions and preventing cyberbullying, hate speech, and harassment. Traditional machine learning methods, such as Support Vector Machines and Naïve Bayes classifiers, have been applied to text classification tasks but often struggle with contextual understanding and generalization [7]. Recent advances in deep learning have shown remarkable progress in sentiment analysis and offensive content detection by leveraging neural networks capable of learning hierarchical textual representations [3], [4]. Hybrid architectures, combining transformer-based models with sequence learning networks like LSTMs, have further improved detection accuracy across varied datasets. The growing diversity of languages and code-mixed data in online spaces has intensified the need for multilingual and robust detection models [4], [5]. Researchers [2] and [6] have demonstrated the effectiveness of transformer-based architectures like BERT in handling multilingual offensive language tasks. Moreover, hybrid deep learning frameworks incorporating attention mechanisms can capture context-sensitive cues and subtle semantic variations that distinguish offensive from non-offensive expressions [3].

This research proposes a model that integrates BERT embeddings, Bidirectional LSTM, and an Attention Mechanism for improved offensive language detection. The model aims to address limitations in existing approaches by combining contextual depth, sequential dependency learning, and interpretability. Experimental validation will be conducted using publicly available datasets such as those on Kaggle [15], ensuring a robust evaluation framework. The proposed hybrid architecture aspires to enhance detection precision, reduce bias in imbalanced datasets, and contribute to safer online environments through intelligent, automated monitoring of harmful textual content.

II. LITRETURE REVIEW

The literature review explores key developments in offensive language detection, focusing on traditional machine learning approaches and their limitations. There is a long history of multilingual text classification and offensive language detection, and we briefly review these aspects in this section.

This work builds a multi-class model for offensive language detection using labeled tweets. After preprocessing and tf-idf feature selection, deep learning models like Bi-GRU, Bi-LSTM, GRU, and Multi-Dense LSTM are applied; Future research should use larger datasets and context-aware models for better generalization. Multimodal approaches and real-time systems can enhance accuracy, while tackling class imbalance, sarcasm, and implicit hate speech. Ensuring fairness and interpretability is

vital for unbiased moderation [1]. This work builds a multi-class model for offensive language detection using labeled tweets. After preprocessing and tf-idf feature selection, deep learning models like Bi-GRU, Bi-LSTM, GRU, and Multi-Dense LSTM are applied. Future research should use larger datasets and context-aware models for better generalization. Multimodal approaches and real-time systems can enhance accuracy, while tackling class imbalance, sarcasm, and implicit hate speech. Ensuring fairness and interpretability is vital for unbiased moderation [2]. Offensive language detection benefits from diverse datasets and advanced models like Transformers that capture subtle context. Multimodal approaches, combining text and images, further enhance detection accuracy in real-world scenarios. Future research should develop real-time systems for social media, while addressing class imbalance, code-switching, sarcasm, and implicit hate speech. Ensuring fairness and interpretability in deployed models will be vital for unbiased moderation [3]. The system performs sentiment analysis and offensive language detection on Tamil-English code-mixed data using ML, DL, and pre-trained models like BERT, RoBERTa, and adapter-BERT. Adapter-BERT achieved the best results with 65% accuracy for sentiment and 79% for offensive detection. Future work can use larger multilingual datasets, improved embeddings, and real-time applications. Transfer learning and fairness across low-resource languages remain key focus areas [4]. The system uses BERT for preprocessing, text representation, and classification into offensive and non-offensive categories. To address multilingualism, both joint-multilingual and translation-based techniques are explored. Experiments on the SOLID bilingual dataset show that the translation-based method with AraBERT achieves 93% F1-score and 91% accuracy. Future research can extend this approach to more languages, larger datasets, and real-time moderation. Enhancing multilingual embeddings and improving adaptability across diverse contexts remain important directions [5].

Dataset Used: Hate Speech and Offensive Language Dataset: The dataset used in this work is "Hate Speech and Offensive Language Dataset developed by Davidson et al. It contains 24,783 labelled tweet's, each manually annotated into three categories Hate Speech, Offensive Language, and Neither. The tweets were collected from Twitter using hate speech-related keywords and then classified through crowd-sourced annotation. Each entry includes the tweet text, class label, and metadata such as tweet ID and user information. This dataset is widely used in NLP research for toxic content detection, sentiment analysis, and social media moderation tasks. It provides a strong benchmark for evaluating machine learning and deep learning models in offensive language detection. However, it also presents challenges like class imbalance, contextual ambiguity, and sarcasm, making it valuable for developing more robust, fair, and interpretable models for online content moderation in research [17].

III. PROPOSED METHODOLOGY

The proposed methodology proposed an efficient hybrid deep learning framework for offensive language detection in textual data. As shown in Figure 3.1, the architecture follows a systematic sequence of steps, beginning with data preprocessing and proceeding through contextual embedding generation, model training, and evaluation. The process starts with a labeled text dataset containing instances of hate, offensive, and neutral language. The text data undergoes pre-processing using the BERT tokenizer, which includes tokenization, normalization, and removal of unnecessary symbols such as URLs and punctuation. This step converts raw text into subword tokens suitable for contextual representation. The BERT model then generates deep contextual embeddings that effectively capture the semantic meaning and relationships between words. These embeddings are fed into a Bidirectional Long Short-Term Memory (BiLSTM) layer to learn sequential dependencies and linguistic patterns in both forward and backward directions. Subsequently, an attention mechanism is applied to assign higher importance to words or phrases that significantly influence the detection of offensive content, enhancing both interpretability and focus. The refined representations are processed through dense layers that perform non-linear transformations to produce the final classification outputs. Finally, the model is evaluated using standard performance metrics such as Accuracy, Precision, Recall, and F1-Score to validate its efficiency. This integrated architecture effectively combines the contextual understanding of BERT, the sequential learning capability of BiLSTM, and the attention mechanism's focus to achieve robust and accurate offensive language detection

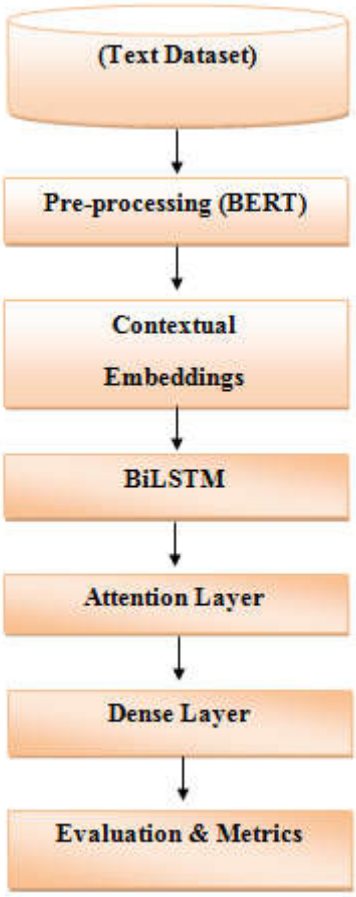

Figure 3.1 : Proposed Architecture Model

IV. RESULT ANALYSIS

The performance evaluation demonstrates that the proposed model consistently outperforms both the baseline BERT model and existing work. Compared to BERT, the proposed framework shows improvements across all metrics, achieving an accuracy of 95.36%, precision of 94.87%, recall of 94.42%, and an F1-score of 94.64%, indicating enhanced classification capability and better handling of complex linguistic patterns. When compared with the study by Simrat Kaur et al. [1], which achieved an accuracy of 91.84% and an F1-score of 89.98%, the proposed model exhibits significant performance gains. These results validate the effectiveness of the proposed methodology, highlighting its robustness, reliability, and suitability for real-world offensive language detection tasks.

Table 4.1: Performance of Proposed Model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT (Baseline)	95.11	92.88	92.24	92.55
Proposed Model	95.36	94.87	94.42	94.64

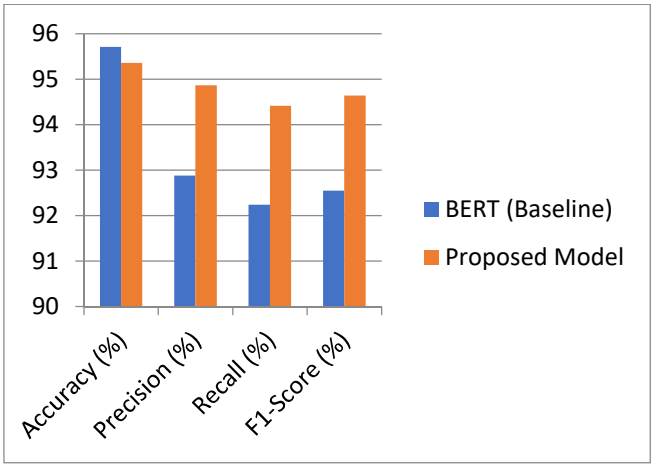

Figure 4.1: Performance of Proposed Model

Table 4.2: Performance Comparison of Proposed Model & existing work

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Simrat Kaur et. al. [R1]	91.84	90.25	89.71	89.98
Proposed Model	95.36	94.87	94.42	94.64

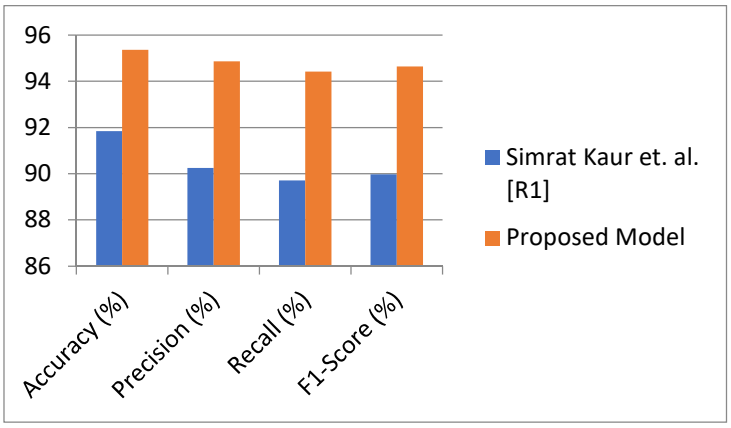

Figure 4.2: Comparative Performance Analysis of the Proposed Model and Existing Approaches

CONCLUSION

The proposed hybrid model effectively advances offensive language detection by leveraging the strengths of BERT embeddings, BiLSTM, and an attention mechanism. Experimental results indicate that it surpasses baseline and existing models, achieving 95.36% accuracy and an F1-score of 94.64%, confirming its robustness and reliability. The attention mechanism enables the model to focus on critical words, improving precision and recall, while the sequence-aware BiLSTM captures contextual flow. This combination ensures better generalization across diverse datasets and resilience against imbalanced data. Overall, the framework provides an efficient, scalable, and accurate solution for moderating harmful content online, contributing to safer digital communication environments and setting a benchmark for future research in offensive language detection.

REFERENCES

[1] Simrat Kaur et. al. “Deep learning-based approaches for abusive content detection and classification for multi-class online user-generated data” International Journal of Cognitive Computing in Engineering Volume

[2] Israe Abdellaoui, et. al. “Investigating Offensive Language Detection in a Low‑Resource Setting with a Robustness Perspective” https://doi.org/10.3390/bdcc8120170, Big Data Cogn. Comput. 2024, https://www.mdpi.com/journal/bdcc

[3] Gulnur Kazbekova et. al. “Offensive Language Detection on Online Social Networks using Hybrid Deep Learning Architecture” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 14, No. 11, 2023

[4] Kogilavani Shanmugavadivel et. al. “Deep learning based sentiment analysis and offensive language identification on multilingual Code-mixed data” https://doi.org/10.1038/s41598-022-26092-3, www.nature.com

[5] Fatima-zahra El-Alami et. al. “A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model” https://doi.org/10.1016/j.jksuci.2021.07.013, science direct 2022

[6] Md Saroar Jahan et al “A systematic review of hate speech automatic detection using natural language processing” Elsevier 2023

[7] Rustam Abdrakhmanov et. al. “Offensive Language Detection on Social Media using Machine Learning” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 15, No. 5, 2024

[8] T. Alsubait and D. Alfageh, “Comparison of machine learning techniques for cyberbullying detection on youtube arabic comments,” International Journal of Computer Science and Network Security, vol. 21, no. 1, pp. 1–5, 2021.

[9] D. Sultan, B. Omarov, Z. Kozhamkulova, G. Kazbekova, L. Alimzhanova et al., "A review of machine learning techniques in cyberbullying detection," Computers, Materials & Continua, vol. 74, no.3, pp. 5625–5640, 2023.

[10] D. Hall, Y. Silva, Y. Wheeler, L. Cheng and K. Baumel, “Harnessing the power of interdisciplinary research with psychology-informed cyberbullying detection models,” International Journal of Bullying Prevention, vol. 4, no.1, pp. 47–54, 2021.

[11] Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M., & Omarov, B. (2021, October). Chatbots and Conversational Agents in Mental Health: A Literature Review. In 2021 21st International Conference on Control, Automation and Systems (ICCAS) (pp. 353-358). IEEE.

[12] T. Ahmed, M. Rahman, S. Nur, A. Islam and D. Das, “Natural language processing and machine learning based cyberbullying detection for Bangla and romanized bangla texts,” TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 20, no. 1 pp. 89–97, 2021.

[13] K.E. Abdelfatah, G. Terejanu, A.A. Alhelbawy, Unsupervised detection of violent content in arabic social media, Comput. Sci. Inf. Technol. (CS IT) (2017)

[14] E.A. Abozinadah, Improved micro-blog classification for detecting abusive arabic twitter accounts, International Journal of Data Mining & Knowledge Management Process (IJDKP) 6 (2016).

[15] https://www.kaggle.com/code/kirollosashraf/hate-speech-and-offensive-language-detection/output

[16] González-Carvajal, S. & Garrido-Merchán, E. C. Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv: 2005. 13012 (2020). 2. Souma, W., Vodenska, I. & Aoyama, H. Enhanced news sentiment analysis using deep learning methods. J. Comput. Soc. Sci. 2(1), 33–46 (2019).