RESEARCH ARTICLE OPEN ACCESS

Offensive-Bot: BERT and NLP based Real-Time Offensive Message Detection and Restriction System

Ms. P. Revathy MCA, M. Phil.*[1], K.Jayashree*[2], R.Preethaa,*[3]
Assistant Professor, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science,Coimbatore, Tamil Nadu, India.
E-mail: revathy@srcas.ac.in
PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India.
E-mail: jayashree2005k@gmail.com
PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India.
E-mail: preethaa9805@gmail.com

-------------------------------------*************************----------------------------------

Abstract:

The rapid expansion of online communication platforms has led to a noticeable rise in offensive language, hate speech, and cyberbullying. Such harmful content affects users emotionally and contributes to unsafe digital spaces. Conventional content moderation methods, including keyword-based filtering and manual supervision, often fail to identify offensive messages that rely on context, implicit meaning, or indirect expressions. To overcome these limitations, this paper presents Offensive-Bot, a real-time offensive message detection system developed using Natural Language Processing (NLP) and a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model. The system examines user messages before they are sent and categorizes them as either safe or offensive. When offensive content is detected, the message is immediately blocked and the user is alerted, encouraging responsible online behavior. In addition, restricted messages are stored for administrative monitoring and future system improvement. Experimental results indicate that the proposed system achieves high detection accuracy with minimal processing delay. Overall, OffensiveBot contributes to creating safer and more respectful online communication environments.

Keywords- Offensive Language Detection, Cyberbullying, Natural Language Processing BERT, Chatbot, Content Moderation, Restriction.

-------------------------------------*************************----------------------------------

I. Introduction

Online communication has become an integral part of daily life through social networking platforms, instant messaging applications, and online discussion forums. These platforms enable fast information exchange and global interaction. However, they also provide opportunities for the spread of offensive language, hate speech, and cyberbullying. Such behavior can cause emotional harm, damage personal reputation, and negatively affect mental health, particularly among younger users.

Most existing content moderation systems rely on keyword-based filtering or manual monitoring by human moderators. Keyword-based approaches are limited because they cannot understand contextual meaning, sarcasm, or indirect forms of abuse. Users can easily bypass these systems by altering spellings or phrasing offensive messages creatively. Manual moderation, while more accurate, is expensive, time-consuming, and impractical for platforms that handle a massive number of messages every day. As a result, harmful content often reaches users before any corrective action is taken.

Recent progress in Artificial Intelligence (AI) and Natural Language Processing (NLP) has enabled more advanced text analysis techniques. Transformer-based models, especially BERT, have demonstrated strong performance in understanding contextual relationships within text. Unlike traditional machine learning methods, BERT analyzes words based on their surrounding context, allowing it to capture the actual intent of a message more accurately.

In this work, an intelligent system named OffensiveBot is proposed to detect and restrict offensive messages in real time. The system aims to prevent harmful content from being transmitted while ensuring smooth and uninterrupted user interaction. By combining contextual text analysis, instant restriction, and administrative monitoring, the proposed approach offers an effective solution for offensive message detection.

1. LITERATURE REVIEW

Several research works have explored offensive language detection, chatbot systems, and transformer-based Natural Language Processing models across different languages and application domains.

Likha et al. (2025) [1] Proposed a comprehensive system for fake news and offensive content detection in the Malayalam language using machine learning, deep learning, and transformer-based models. Their study evaluated multilingual transformer models such as mBERT, IndicBERT, XLM-RoBERTa, and MuRIL along with traditional classifiers. Explainable AI techniques including LIME and Anchor were used to interpret model predictions. The results showed that transformer-based models achieved very high accuracy, demonstrating their effectiveness in handling complex and code-mixed language content.

Attigeri and Agrawal [2] developed an NLP-based chatbot system to assist students and parents during the engineering college admission counseling process. Their work compared TF-IDF, pattern matching, and neural network-based models. The sequential neural network model achieved superior accuracy by effectively understanding contextual information in user queries. This study highlights the importance of contextual modeling and real-time response generation in chatbot-based systems.

Sreelakshmi et al. [3] focused on detecting hate speech and offensive language in CodeMix Dravidian languages using a cost-sensitive learning approach. The authors combined transformer-based embeddings such as MuRIL and BERT with SVM classifiers to handle class imbalance issues. Experimental results showed improved accuracy across Malayalam-English, Tamil-English, and Kannada-English datasets, proving the robustness of cost-sensitive learning for low-resource languages.

Cao et al. (2024) [4] introduced a dual-channel offensive language detection model for Chinese text by integrating BERT-based semantic representations with topic-level information generated using Correlated Topic Models. The proposed approach employed parallel CNN architectures and multi-head attention mechanisms to capture both local and global text features. The model achieved high accuracy and demonstrated strong performance in understanding nuanced offensive content. **Molero et al. (2023) [5]** evaluated offensive language detection methods for Spanish social media text by comparing classical Bag-of-Words approaches with transformer-based models. Their findings showed that transformer models significantly outperformed traditional methods, particularly when trained on domain-specific social media data. The study emphasized the importance of handling class imbalance and appropriate preprocessing techniques for real-world datasets.

3.METHODOLOGY

The proposed methodology focuses on identifying and restricting offensive messages instantly using Natural Language Processing and a BERT-based classification model. The system workflow consists of multiple stages designed to ensure reliable and efficient detection.

3.1. Data Collection

Textual data containing both offensive and non-offensive messages is collected from publicly available sources and online communication datasets. The collected messages are carefully labeled to support effective training and evaluation of the model.

3.2. Text Preprocessing

Before classification, the collected text is cleaned to remove unwanted symbols and inconsistencies. The text is converted into a standardized format, and tokenization is applied to prepare the input for the BERT model.

3.3. Contextual Representation Using BERT

The preprocessed text is transformed into contextual embeddings using the BERT tokenizer. This representation allows the model to understand the meaning of words based on their surrounding context rather than treating them independently.

3.4. Model Training and Classification

The BERT model is fine-tuned using the labeled dataset to distinguish between offensive and safe messages. During training, the model learns patterns commonly associated with offensive language. Once trained, the model is capable of classifying incoming messages in real time.

3.5. Real-Time Message Filtering

When a user submits a message, it is immediately analyzed by the trained model. Safe messages are delivered normally, while offensive messages are blocked and a warning is displayed to the user.

4.PROPOSED MODEL

The proposed system, Offensive-Bot, is implemented as an intelligent chatbot that integrates NLP techniques with a BERT-based offensive message detection mechanism. The model ensures proactive prevention of harmful content while maintaining a seamless user experience.

4.1 System Components

ystem Components

- **User Interface**

Provides a chat environment for users to enter messages, which are analyzed before transmission.

- **NLP Processing Module**

Handles text cleaning, tokenization, and conversion into BERT-compatible input.

- **BERT Classification Engine**

Analyzes contextual meaning and classifies messages as offensive or non-offensive.

Restriction and Alert Module

Blocks offensive messages and immediately notifies users.

- **Admin Monitoring Module**

Stores restricted messages for monitoring, analysis, and future improvements.

4.2 Data Flow

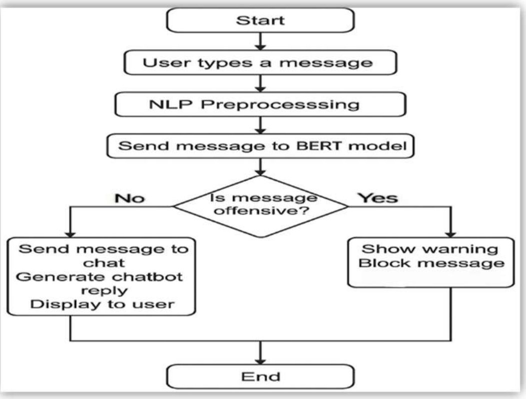


Fig 1 Data flow of the model

1. The user enters a message through the chat interface.
2. The message is forwarded to the NLP preprocessing module.
3. The processed message is analyzed by the BERT classifier.
4. Based on the result, the message is either delivered or restricted.
5. Restricted messages are logged for administrative review.

4.3 System Deployment

The system is deployed as a web-based application connected to a backend server. The BERT model operates on the server to handle message classification efficiently. The architecture supports scalability, allowing multiple users to interact with the system simultaneously with minimal latency.

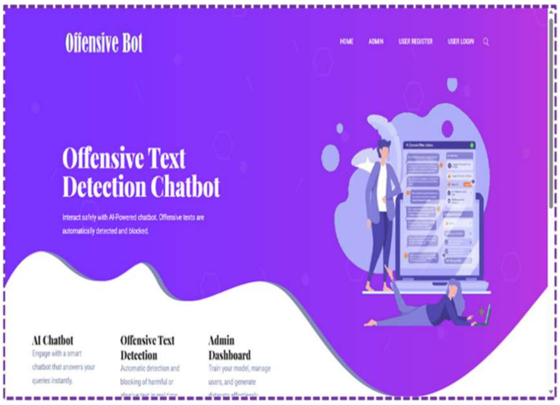
4.4 System Benefits

- Accurate detection of offensive and abusive language
- Real-time message restriction before delivery
- Enhanced online safety and user experience
- Reduced reliance on manual moderation
- Scalable and adaptable system architecture

5.RESULTS AND DISCUSSION

The performance of Offensive-bot was evaluated based on detection accuracy, response time, and overall user experience. The BERT-based classifier demonstrated strong performance in identifying offensive messages, including indirect and context-dependent abuse. Compared to keyword-based systems, the proposed approach significantly reduced false detections.

The system processed messages with minimal delay, ensuring real-time restriction of harmful content. Immediate user alerts encouraged more responsible communication behavior. Additionally, logging restricted messages enabled effective administrative monitoring and contributed to system reliability. Overall, the results indicate that the proposed system successfully maintains a safer online communication environment.



Offensive Text Detection Chatbot

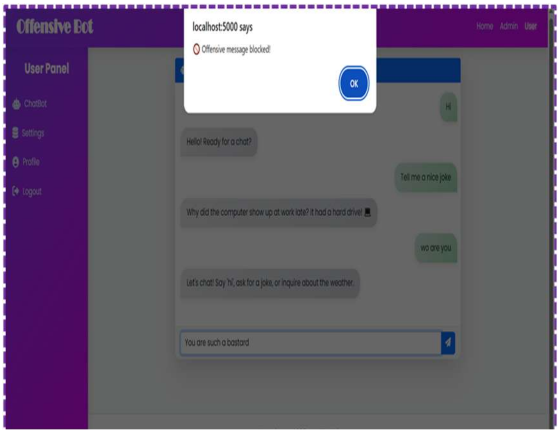
Interact safely with AI-Powered chatbot. Offensive texts are automatically detected and blocked.

AI Chatbot
Engage with a smart chatbot that answers your queries instantly.

Offensive Text Detection
Automatic detection and blocking of harmful or abusive text in real-time

Admin Dashboard
Train your model, manage users, and generate datasets effortlessly

Fig: 2 Application Home page



localhost:5000 says

Offensive message blocked!

Hi

Hello! Ready for a chat?

Tell me a nice joke

Why did the computer show up at work late? It had a hard drive! 💻

wo are you

Let's chat! Say 'hi', ask for a joke, or inquire about the weather.

You are such a bastard

Fig:3 Blocking of Offensive messages

6.CONCLUSION

This paper presented Offensive-Bot, a real-time offensive message detection system developed using Natural Language Processing and a BERT-based classification model. The system effectively addresses the limitations of traditional moderation techniques by understanding contextual meaning and preventing offensive messages before transmission.

By integrating real-time restriction, user notification, and administrative monitoring, the proposed system enhances online safety and promotes responsible digital communication. Future enhancements may include multilingual support, voice-based message analysis, and multimodal content detection. Continuous model training with real-world data can further improve detection accuracy and adaptability. Overall, OffensiveBot serves as a scalable and intelligent solution for modern online communication platforms.

REFERENCES

[1]. J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. NAACL*, 2019.

[2] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection," *Proc. NAACL*, 2016.

[3] P. Fortuna and S. Nunes, “A Survey on Automatic Detection of Hate Speech in Text,” *ACM Computing Surveys*, 2018.
[4] T. Davidson et al., “Automated Hate Speech Detection and the Problem of Offensive Language,” *Proc. ICWSM*, 2017.
[5] S. Minaee et al., “Deep Learning Based Text Classification: A Comprehensive Review,” *IEEE Access*, 2021.

[6] Zhang, Y., & Wallace, B. “A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification” arXiv preprint arXiv:1510.03820.,2017.
[7] Goyal, P., Gupta, R., & Goyal, L. M, “A review of chatbot and natural language processing. International Journal of Advanced Research in Computer Science, 11(4), 69-75.2020
[8] Rashid, S. M., Abdullah, A. H., & Ahmed, M. A. “Development of a chatbot using natural language processing for customer service. International Journal of Computer Science and Information Security (IJCSIS), 17(5), 167.-2019
[9] Lowe, R., & Pow, N, “The rise of the conversational interface: A new kid on the block. Computer”, 50(8), 58-63,2017.
[10] Rajabi, A., Asgarian, A., & Ebrahimi, M. “A comparative study of machine learning algorithms for automated response selection in chatbot systems. In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 45-52), 2018.