IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 3, March 2025

DOI: 10.17148/IJARCCE.2025.14344

Harmful Content Detection on Social Media Platforms

Dr. B. Sivaranjani[1], Ms. M. Divyadharshini[2], Ms. L. Glory[3]

Professor Department of Computer Science, Department of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore[1]

B. Sc Computer Science, Department of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore[2,3]

Abstract: This project "Harmful Content Detection on Social Media Platform "is Python based project. It is designed using PYTHON/FLASK as front end and PHP as backend. The web application for the detection of offensive word is used to find the offensive Word and it performs a comparison between the words stored in the database and the words present in the text. The system then shows the user if any offensive words are detected. It shows the offensive and non-offensive words in graphical representation like chart, bar graph to find the presence of offensive word in the text. The proposed system is tested on a dataset of offensive words, and the results show that it can effectively detect offensive words in offline mode. Harmful Content Detection On Social Media Platforms implements our coded, machine learning algorithms, in finding a negative comment from the messages it receives by a user. The algorithm first gives the message a value and then based on our pre trained data, it decides if the comment is harsh enough to be transformed or not. It is assigned a value and if the value results in a positive sentence, the system will proceed to send the transformed positive sentence to the end user. Otherwise, the sentence will be placed through the models again. The users communicate through a developed web front face and they are connected to a central server. The users are termed as clients. If any messages are modified the receiving user will be notified along with the modified message. A major source of cyberbullying is social media. These platforms can have the opposite desired effect of uniting peers, and instead can be weaponized to harass and bully others. Most existing solutions have shown techniques/approaches to detect cyberbullying, but they are not freely available for end-users to use. They haven't considered the evolution of language which makes a big impact on cyberbullying text. It doesn't affect only for health, there are more different aspects which will lead life to a threat. Cyberbullying is a worldwide modern phenomenon which humans cannot avoid hundred percent but can be prevented.

Keywords: Harmful Content Detection, Social Media Moderation, Offensive Word Detection, Cyberbullying Prevention, Machine Learning Algorithms, Sentiment Analysis, Natural Language Processing (NLP), Flask Web Application, PHP Backend, Text Classification, Data Filtering, Content Moderation, Automated Censorship.

I.INTRODUCTION

The aim of this project is to develop a Python-based program that detects offensive words in a given text. The program should be able to read text input, identify words that are potentially offensive, and display them accordingly. The system is designed to function effectively in offline mode. The overall goal of this project is to develop a useful tool for individuals and businesses who want to ensure that their communication remains free of offensive language.

Despite its many benefits, social media also poses significant risks. While it helps establish new relationships and maintain existing friendships, it also increases the risk of children and users being confronted with threatening situations, including grooming, sexually transgressive behavior, cyberbullying, and exposure to depressive and suicidal content. This project aims to mitigate these risks by implementing an offensive word detection system.

The project "Harmful Content Detection on Social Media Platform" aims to create a web application using Python Flask that will detect offensive words in user-generated text. The system will analyze the input text and identify any offensive words present by comparing them with words stored in a database.

Harmful content detection on social media platforms is achieved through the implementation of our coded machine learning algorithms, which are used to analyze user messages.The algorithm first assigns a numerical value to the message and then, based on pre-trained data, determines whether the comment is offensive. If the comment is deemed offensive, the system will transform it into a more neutral or positive statement before displaying it to the receiving user. If a message is modified, the receiving user will be notified and provided with the modified version.

DOI: 10.17148/IJARCCE.2025.14344

The application will feature a user-friendly interface, allowing users to input text and receive feedback regarding offensive words in real-time. To enhance the user experience, the system will present a graphical representation of offensive and non-offensive word distributions through bar charts and graphs. This enables users to visualize the prevalence of offensive language in their text submissions.

The primary technologies used in this project include:

- Front-end: Python (Flask Framework)
- Back-end: PHP
- Database: Predefined dataset of offensive words
- Machine Learning: NLP-based sentiment analysis and classification models

The objectives of this project are

- To develop an efficient tool that detects offensive words in text-based communication.
- To provide an intuitive web-based interface for users to check text for offensive content.
- To implement machine learning algorithms for content classification and modification.
- To notify users when a message is modified due to offensive content.
- To help prevent cyberbullying and ensure safer online communication.

Cyberbullying is a growing global issue, affecting individuals of all ages. Many existing solutions detect offensive content but are not freely available for public use. Moreover, they often fail to consider the evolution of language and slang, which plays a significant role in cyberbullying text. This project aims to provide an accessible solution that not only detects harmful content but also transforms negative comments into more positive alternatives. By implementing this system, individuals and organizations can foster healthier communication environments on social media platforms.

II.LITERATURE REVIEW

Harmful content detection has emerged as a critical area of research, particularly with the rapid growth of social media platforms. Online communication allows individuals to connect worldwide; however, it also exposes users to offensive language, cyberbullying, and harmful behavior. Developing robust detection systems requires a combination of Natural Language Processing (NLP), Machine Learning (ML), and Sentiment Analysis to identify and mitigate harmful content effectively. Several studies have proposed techniques for harmful content detection, focusing on keyword matching, machine learning classifiers, and context-based analysis. However, achieving high accuracy while ensuring minimal false positives remains a significant challenge.

III.RESEARCH ON HARMFUL CONTENT DETECTION

3.1 Keyword-Based Detection Systems

Keyword matching techniques were among the earliest methods for detecting offensive content. These systems rely on predefined lists of offensive words or phrases.

- Cheng et al. (2019) developed a keyword-matching algorithm that flags offensive language in text-based communication. While effective for detecting explicit words, this approach struggles with sarcasm, slang, and evolving language patterns.
- Davidson et al. (2017**)** improved keyword-based detection by integrating contextual analysis to reduce false positives. This method enhanced accuracy but required constant updates to the keyword database.
  Limitations:
- Limited adaptability to new slang and informal language.
- High false positive rates in cases of ambiguous text.

3.2 Machine Learning-Based Approaches

Machine learning models have shown significant improvement in identifying harmful content through pattern recognition and contextual understanding.

- Park et al. (2018) applied a Support Vector Machine (SVM) model trained on labelled offensive text data. The model achieved high precision in identifying direct offensive comments but struggled with implicit harmful content.
- Zhang et al. (2020) used a Convolutional Neural Network (CNN) for text classification. The CNN model effectively identified offensive language in short social media comments but required large-scale labelled datasets.

DOI: 10.17148/IJARCCE.2025.14344

- Badjatiya et al. (2017) implemented a Recurrent Neural Network (RNN) model that integrated word embeddings to understand contextual word meanings. This improved detection accuracy for subtle and indirect harmful content.

Limitations:

- ML models heavily depend on labeled data for training.
- Contextual nuances like sarcasm and double meanings pose challenges.

3.3 Sentiment Analysis in Harmful Content Detection

Sentiment analysis techniques are widely used to classify text based on emotional tone. These models analyze language to identify positive, neutral, or negative sentiment.

- Rashid et al. (2021) implemented a Bidirectional Long Short-Term Memory (Bi-LSTM) model to improve sentiment-based detection. This model effectively classified toxic language by analysing word sequences and emotional context.
- Kumar et al. (2020) combined sentiment analysis with TF-IDF (Term Frequency-Inverse Document Frequency) for keyword weighting, improving offensive content detection in multilingual datasets.

Limitations:

- Sentiment analysis models may misinterpret humor, sarcasm, or culturally specific phrases.

3.4 NLP-Based Techniques

Natural Language Processing (NLP) techniques provide advanced solutions for harmful content detection. These methods analyze text structure, word relationships, and contextual meanings.

- Devlin et al. (2018) introduced BERT (Bidirectional Encoder Representations from Transformers), which significantly improved NLP performance in content detection by understanding contextual word relationships.
- Liu et al. (2019) combined BERT with Word2Vec embeddings to create a hybrid model that effectively identified harmful content in tweets and online discussions.
- Kumar et al. (2023) introduced GPT models for improved offensive language detection by leveraging large-scale text datasets for adaptive learning.

Limitations:

- NLP models require extensive computational resources.
- Complex sentence structures, sarcasm, and evolving slang remain challenging.

3.5 Real-Time Content Moderation Systems

Some advanced frameworks integrate ML models with web applications for real-time detection and prevention.

- Facebook's Deep Text (2016) uses deep learning to understand text content in multiple languages, enhancing harmful content detection.

Twitter's Birdwatch (2021) employs community-driven reporting systems combined with NLP models to improve content moderation.

Limitations:

- Real-time detection systems often struggle with scalability and false positive issues.
- Detecting multimedia content (images/videos with offensive context) remains a growing challenge.
- Limited adaptability to evolving slang and dynamic language patterns.
- Difficulty in detecting subtle threats like sarcasm, humor, or coded language.
- Existing solutions often focus on detection alone but lack mechanisms for positive content transformation.
- Many available tools are proprietary, making them inaccessible to small businesses and educational institutions.

IV.PROPOSED SOLUTION ENHANCEMENT

The proposed system addresses these gaps by:

- Combining machine learning, NLP, and sentiment analysis to detect and transform offensive language into neutral or positive content.
- Implementing a user notification system that alerts recipients when a message is modified, improving transparency.
- Enhancing user experience through graphical visualizations of offensive word frequency for better understanding.
- Developing a Flask-based interface integrated with PHP for efficient data handling and interactive results

DOI: 10.17148/IJARCCE.2025.14344

V.CONCLUSION

This literature review highlights the strengths and limitations of existing harmful content detection systems. While traditional approaches focus on detection alone, this project introduces a transformative solution that enhances text positivity while maintaining communication flow. By integrating dynamic visual feedback and transparent user notifications, the proposed system aims to create a safer and healthier online environment.

REFERENCES

[1] A. Mayrock, The Survival Guide to Bullying. New York, NY, USA: Scholastic, 2015.

[2] D. A. Smith and S. M. Mohammad, Automated Hate Speech Detection and the Problem of Offensive Language. Cambridge, MA, USA: MIT Press, 2018.

[3] M. W-Band and M. A. Chenoweth, Text Analysis and Its Applications. London, UK: Springer, 2020.

[4] S. Goyal, Automatic Detection of Hate Speech in Social Media. Berlin, Germany: Springer, 2021.

[5] A. Fortuna and S. Nunes, “A survey on automatic detection of hate speech in text,” arXiv preprint, 2018. [Online]. Available: https://arxiv.org/pdf/1810.04808.pdf

[6] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” Neural Computing and Applications, vol. 29, no. 7, pp. 1–10, 2017. [Online]. Available: https://link.springer.com/article/10.1007/s00521-017-3271-x

[7] M. A. Mohammad, “Random forest classifier for offensive text detection,” ResearchGate, 2017. [Online]. Available: https://www.researchgate.net/publication/320959427_Random_Forest_Classifier_for_Offensive_Text_Detection

[8] ABA Legal Fact Check, “Hate speech and the First Amendment,” 2018. [Online]. Available: https://abalegalfactcheck.com/articles/hate-speech.html