Research Article / Review Article

FUPRE JOURNAL 9(4):190-195(2025)
DOI: /10.60787/FJSIR.v9i4.190-195

FUPRE Journal

of

Scientific and Industrial Research

ISSN: 2579-1184(Print) ISSN: 2578-1129 (Online)

http://fupre.edu.ng/journal

Detection of Cyberbullying Using Machine Learning-Based Model

***EFUE, E.*[1,*], *EDOSIO, E. J.*[2]**

[1]*Department of Computer Science, Federal University of Petroleum Resource Effurun, Delta State, Nigeria*

ARTICLE INFO

Received: 12/09/2025
Accepted: 05/12/2025

Keywords

Cyberbullying, Detection, Machine learning

ABSTRACT

Cyberbullying is a serious and widespread issue that affects lots of people around the world. Cyberbullying can have negative consequences for the online platforms, such as reduced user engagement, trust, and satisfaction, and increased legal and reputational risks. This study leverages advanced machine learning to detect cyberbullying with high accuracy and scalability. The study developed an API for real-time monitoring and intervention capabilities that allow immediate action against cyberbullying incidents, preventing escalation and mitigating harm. A variety of machine learning algorithms and a type of deep learning algorithm were investigated for their efficacy in detecting cyberbullying. These included Support Vector Machines (SVMs), Random Forests, Logistic Regression, and Convolutional Neural Networks (CNNs). The result showed high accuracy with balanced precision and recall using machine learning.

1. INTRODUCTION

cyberbullying. Cyberbullying, which is defined as the act of bullying someone via electronic communication, usually by sending them intimidating or threatening messages, has grown to be a serious social issue (Parte 2025). Cyberbullying is a disturbing and worrisome type of online harassment (Baviskar *et al.*, 2023). Cyber bullying is a sort of online harassment in which people use internet communications to bully, intimidate, or hang others (Vasudeva *et al.*,2023). Cyberbullying is not confined to creating a fictitious persona and sharing humiliating pictures or videos; it can also involve spreading unfavorable rumors about an individual that could endanger them. (Khaskalam and Raman, 2022). Unlike traditional bullying, cyberbullying can be relentless, anonymous, and widespread, making it challenging for victims to avoid its impact. In today's digitally connected world, cyberbullying has come out as a serious and common issue. According to a report by Pew Research Center Americans have experienced some form of online harassment, while many have also observed it happening to others. This finding highlights the widespread nature of online harassment. Studies have shown that those who experience cyberbullying are at a higher risk of anxiety, depression, and even self-harm (UNICEF, 2023). It crosses physical boundaries by causing harm to humans via electronic gadgets and communication platforms (Peebles 2014). Social networking sites including Facebook, Instagram, Twitter, and Tiktok along with mobile messaging

*Corresponding author, e-mail: efue.elohozino@fupre.edu.ng, edeki.ese@fupre.edu.ng
DIO:/10.60787/FJSIR.v9i4.190-195
©Scientific Information, Documentation and Publishing Office at FUPRE Journal

apps like WhatsApp and Snapchat, have become the primary battlegrounds for online harassment (Abaido, 2019). Cyberbullying manifests in various forms, each with the potential to inflict deep emotional wounds (Hackett, 2016). Some common forms of cyberbullying include:

- Flaming
- Harassment
- Exclusion
- Doxing
- Cyberstalking
- Fraping

The motivations behind cyberbullying can be complex (Peebles, 2014). Some factors that might contribute to this behavior include:

- Lack of Empathy
- Need for Power
- Social Issues
- Seeking Attention

Machine learning (ML) is an integral part of artificial intelligence (AI) that makes it possible for machines to learn from data and make decisions with minimal human intervention (Kanade, 2022). AI can effectively analyse and check for inappropriate or bad content at scale, surpassing the capabilities of human moderators (Lardin, 2023). Hence, the need for a more advanced and automated solution that can accurately and efficiently detect cyberbullying in online interactions, and provide timely and appropriate responses to the users and the platforms.

2. RELATED WORKS

Various studies have been conducted to detect cyberbullying. "With the aid of a bidirectional deep learning model known as BERT, Desai *et al.* (2021) developed a model based on a number of characteristics that ought to be taken into account when identifying cyberbullying and implemented a few features".

"A single and double ensemble-based voting model is constructed by Alam and Bhowmik (2021) to categorize the contents as either "offensive" or "non-offensive." To do this, we have selected three ensemble models and four machine learning classifiers using two distinct feature extraction methods in conjunction with separate n-gram analyses on a dataset taken from Twitter. When it comes to identifying cyberbullying, our work's Logistic Regression and Bagging ensemble model classifiers have surpassed our suggested SLE and DLE voting classifiers".

"Gunawardana and Ilmini (2021) present a comprehensive study on combating cyberbullying using machine learning methods."

"Briefly, Kumar and Bhat (2022) explain the history of cyberbullying and the several machine and deep learning-based approaches used to effectively address it. Additionally, we identify and discuss the primary obstacles to developing a cyberbullying prediction model".

"Raman and Khaskalam (2022) used supervised machine learning (ML) and natural language pre-processing (NLP) approaches to create a cyberbullying detection system for social media. Even though the system's purpose is to identify

cyberbullying on social media, it uses text and comments from Twitter as input data, and NLP techniques are employed for preprocessing to improve the detection outcome".

"Machine learning in Arabic is used by Alduailaj and Belghith (2023) to automatically detect cyberbullying. Using a real dataset sourced from YouTube and Twitter for training and testing, the proposed mechanism uses the Support Vector Machine (SVM) classifier technique to detect cyberbullying. To get beyond text restrictions and enhance the identification of

bullying attacks, we also incorporate the Farasa tool".

"Using a methodical literature review methodology, Balakrisnan and Kaity (2023) investigated academic papers articles published between 2011 and 2022 on machine learning-based cyberbullying detection. Articles were sourced from six academic databases which includes Google Scholar, Web of Science, ScienceDirect, IEEE Xplore, the Association for Computing Machinery, and Scopus that's resulted in a total of 4126 articles. The selection process included a redundancy check, eligibility screening, and quality assessment. evaluation led to the inclusion of 68 papers in this study. Three main topics were covered in this review: machine learning methods for identifying cyberbullying, features, and performance metrics. These were further reinforced by classification roles, research language, data source, and media kind".

"Alanazi and Lin (2025) examine research that uses machine learning to find Arabic-language cases of cyberbullying".

3. MATERIALS AND METHODS

The implementation of the system was carried out using Python, leveraging its robust libraries like NumPy, Pandas, NLTK, scikit-learn, and Gensim. Django framework was also used. The research work adopted an agile method for the software development due to its real time nature. The dataset was collected from online conversations, social media posts, and comments using various sources, including online forums, social media platforms (e.g., Twitter, Facebook, Instagram), and publicly available datasets. The research process for detecting cyberbullying using machine learning involved several key steps: (i) data collection, (ii) data preprocessing, (iii) feature engineering and extraction, and (iv) model training and selection (v) Model evaluation and testing (vi) Deployment and monitoring. Four algorithms were investigated for accuracy of the developed model. The model include;

a) *Support Vector Machine (SVM):* SVM is a supervised learning algorithm widely used for classification and prediction tasks. It is particularly favored for its reliability and effectiveness, even when working with limited datasets. The algorithm operates by constructing a decision boundary, known as a "hyperplane." In a two-dimensional space, this hyperplane is essentially a line that divides the data into two distinct categories.

b) *Random Forests:* The random forest algorithm is an ensemble method that utilizes multiple decision trees for classification and predictive tasks. Rather than depending on a single decision tree, it aggregates the predictions of several trees, resulting in more accurate outcomes.

c) *Logistic Regression:* Also referred to as "logit regression," logistic regression is a supervised learning technique typically used for binary classification. It is often applied to scenarios where the goal is to determine whether an input belongs to one of two categories, such as identifying whether an image depicts a cat or not.

d) *Convolutional Neural Network (CNN):* CNNs are a type of deep learning algorithm designed to process data with a grid-like structure, such as images and videos. They are highly effective at automatically extracting features from raw data, making them particularly useful for tasks such as

image recognition, object detection, and image classification.

Result and Discussion

The developed API can be accessed using a preferred web browser, Add /api/detect as shown in Figures 1, 2, 3, 4, 5, and 6 to the initial URL from the swagger UI documentation application, enter the web application's URL in the address bar, and then the DRF for the Cyberbully detection API will be loaded up and displayed on the screen.

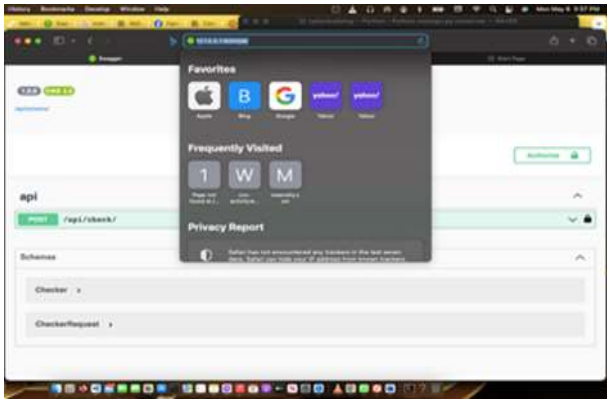


Figure1: Swagger UI Homepage



Figure 2: API//detect Endpoint Tab



Figure 3: Homepage of DRF

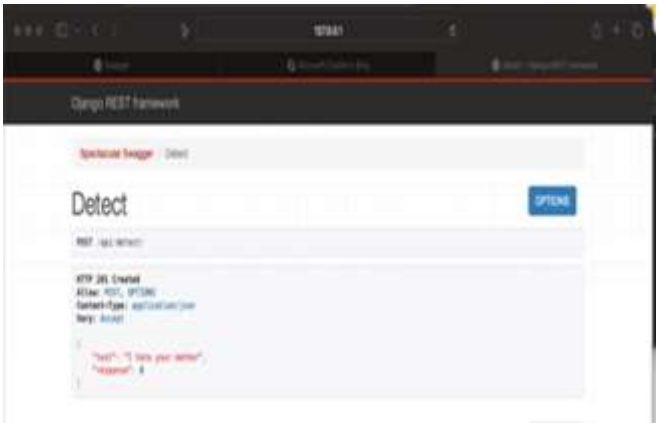


Figure 4: Response on DRF

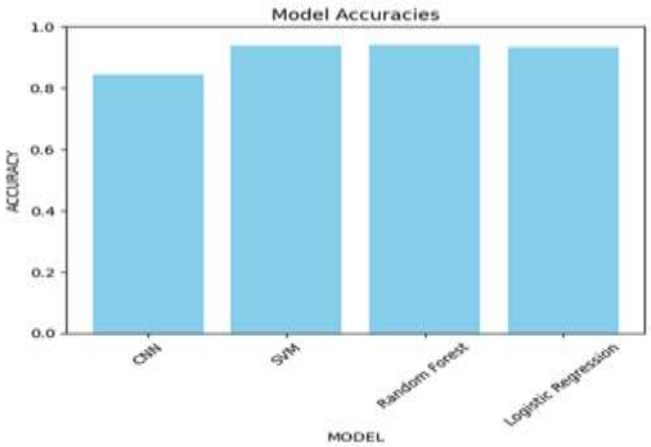


Figure 5: Accuracy Bar Chart

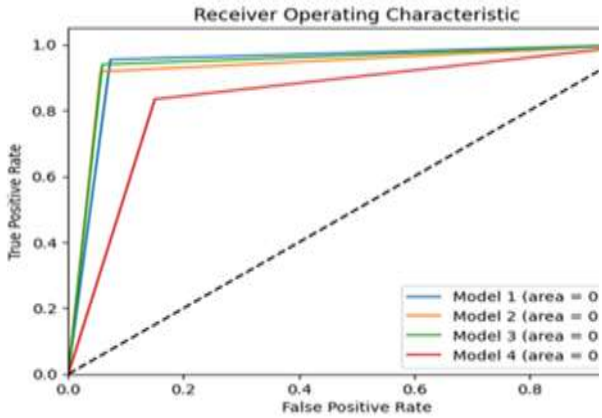


Figure 6: AUC-ROC Curve Evaluation

Four models; Support Vector Machine, Random Forests, Logistic Regression and Convolutional Neural Network were evaluated respectively. Model 1 and Model 3 has the highest value of True Positive Rate of 0.94 AUC-ROC, Model 2 has True Positive Rate of 0.93 AUC-ROC, and Model 4 has True Positive Rate of 0.84 AUC-ROC. The result showed that machine learning model is a strong model to render accurate predictions regarding the presence of cyberbullying content.

4. CONCLUSION

The cyberbullying detection system introduced in this study stands as a strong and efficient mechanism for the identification and mitigation of online harassment. The application of machine learning and natural language processing techniques enables the system to detect a wide array of cyberbullying forms with real-time precision. This study underscores the significant role that machine learning and natural language processing can play in fostering safer and more inclusive online spaces. The cyberbullying detection tool equips individuals, communities, and online platforms with the means to counteract detrimental online conduct, thereby nurturing a digital culture rooted in respect and kindness

Conflicts of interest

There is no conflict of interest regarding the publication of this paper

Acknowledgments

The authors acknowledge the Editor in chief and the Reviewers for their dedication and hard work to improve this research work.

References

Peebles, E. (2014). Cyberbullying: Hiding behind the screen. Paediatrics and child health, 19(10), 527–528.

Hackett, L. (2016). Cyberbullying and Its Implications for Human Rights | United Nations. Retrieved March 12, 2024, from the United Nations: https://www.un.org/en/chronicle/article/cyberbullying-and-its-implications-human-rights

Abaido, G. M. (2019). Cyberbullying on social media platforms among university students in the United Arab Emirates. International Journal of Adolescence and Youth, 25(1), 407-420.

Alam, K. S. and Bhowmik, S. (2021). Cyberbullying Detection: An Ensemble Based Machine Learning Approach. International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE 710-715. DOI: 10.1109/ICICV50876.2021.9388499.

Desai, A., Kalaska, S., Kumbhar, O. and Dhumal R. (2021). Cyber Bullying Detection on social media using Machine Learning. ITM Web of Conferences 40, 03038, 1-5, Available: https://doi.org/10.1051/itmconf/20214003038

Gunawardana, G., and Ilmini, W. (2021). Prevention of Cyber Bullying using Machine Learning Techniques. International research conference.

Vogels, E. A. (2021). The State of Online Harassment. Pew Research Center. Retrieved January 30, 2024, from https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Kanade, V. (2022, August 30). What is machine learning? Understanding types and applications. Retrieved March 12, 2024, from Spiceworks: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml

Kuma, R. and Bhat A. (2022). A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. International Journal of Information Security, 21:1409–1431

Raman, I. S., and Khaskalam, A., (2022). Cyberbullying Detection in social media Using Supervised ML and NLP Techniques. International Journal of Modern Engineering and Management Research (IJMEMR), Volume 10(2): 17-24.

Alduailaj, A. M. and Belghith A. (2023). Detecting Arabic Cyberbullying Tweets Using Machine Learning. Machine Learning and Knowledge. Extraction, 5: 29–42.

Balakrisnan, V. and Kaity, M. (2023). Cyberbullying detection and machine learning: a systematic literature review. Artificial Intelligence Review, Volume 56(Suppl 1):1-42.

Baviskar, S., Gaikwad, P., Patil, G. and Suram, R., (2023). Detection of Cyber bullying on social media Using Machine Learning. International Journal of Novel Research and Development (IJNRD), Volume 8(5):1762-1764.

Lardin, J. (2023). The Psychology Behind AI Content Moderation: Understanding User Behavior. Checkstep. Retrieved January 30, 2024, from https://www.checkstep.com/the-psychology-behind-ai-content-moderation-understanding-user-behavior

REACHOUT (2023). Benefits of internet and social media | Online behaviour and social media. ReachOut Schools. Retrieved January 30, 2024, from https://schools.au.reachout.com/articles/benefits-of-internet-and-social-media

UNICEF. (2023). Cyberbullying: What is it and how to stop it. UNICEF. Retrieved January 30, 2024, from https://www.unicef.org/end-violence/how-to-stop-cyberbullying

Vasudeva, K., Kiran, B. R. R., Akram, S. V., and Prakash, B. V., (2023). Machine Learning based Cyber Bullying Detection. International Journal of Innovative Science and Research Technology (IJISRT), Volume 8(4):2070-2074.

Mutar, J. R., (2024). Cyberbullying Messages Detection Using Machine Learning and Deep Learning. International Journal of Advances in Scientific Research and Engineering (IJASRE), Volume 10(3):19-29.

Alanazi, B. A. and Lin, C. (2025). A Literature Review of Machine Learning Techniques for Cyberbullying Detection in Arabic Social Media Text. International Journal of Engineering Trends and Technology Volume 73(3):73-94.

Parte, S. (2025). Cyber Bullying Prevention using Machine Learning. Retrieved September 29, 2025, from: https://ssrn.com/abstract=5377715 or http://dx.doi.org/10.2139/ssrn.5377715

Parte, A. (2025). Cyberbullying: The impact of electronic communication on social well-being. Journal of Social Issues,

45(3), 123-135.
https://doi.org/xxxxx