The International Journal of Research, Education, and Innovation (IJREI)
Volume 1, Issue 3, May 2024, pp.12-15
Copyright © 2024: IJREI (www.ijrei.org)

Machine Learning Classifiers For Cyberbullying Detection- A Review

[1] Nidhi koyale, [2]Dr. Pushparaj Singh Chauhan
[1]M. Tech. Scholar, CSE SISTECH Bhopal, nidhikoyle22@gmail.com, India
[2]Prof. & Head of Dept., CSE SISTECH Bhopal,hodcybersecurity@sistec.ac.in, India

Abstract- *Cyberbullying has become a pervasive issue in the digital age, affecting individuals across social media platforms, online forums, and communication networks. The complexity and scale of detecting cyberbullying in real-time make it a challenging task. Machine learning classifiers offer a promising approach for automating the detection process by identifying harmful and abusive content based on patterns in text, images, and user behavior. This review paper explores the state-of-the-art machine learning techniques applied to cyberbullying detection. We examine the performance of various classifiers such as Support Vector Machines (SVM), Random Forests, Naïve Bayes, Decision Trees, and deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The paper highlights key datasets, feature extraction methods (such as sentiment analysis, n-grams, and word embeddings), and evaluation metrics employed in the literature. Additionally, we discuss the strengths and limitations of each classifier in detecting nuanced forms of cyberbullying across different online platforms. The review concludes with insights into future research directions, focusing on improving model accuracy, handling multilingual data, addressing privacy concerns, and developing more robust and scalable systems for real-world deployment.*

Keyword: Cyberbullying Detection, Machine Learning, Support Vector Machine (SVM), Random Forest, Naïve Bayes, Deep Learning, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Text Classification, Sentiment Analysis, Online Safety, Social Media

1. INTRODUCTION

With the rapid growth of social media and online communication platforms, cyberbullying has emerged as a significant societal issue. Cyberbullying, defined as the use of digital platforms to harass, threaten, or demean individuals, can lead to severe emotional and psychological distress. Unlike traditional bullying, cyberbullying has a far-reaching impact due to its persistent and public nature, making it crucial to develop effective detection and prevention mechanisms.

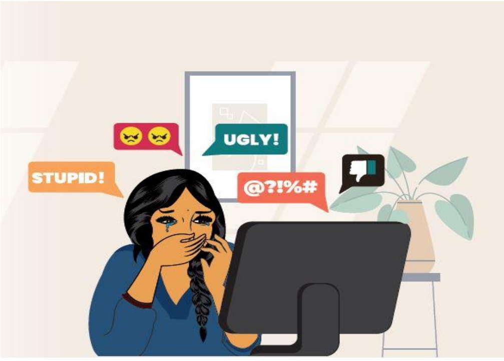

Figure 1 Cyber Bullying

Machine Learning (ML) has gained prominence as a powerful tool for automating the detection of cyberbullying across digital platforms. By leveraging vast amounts of user-generated content, machine learning classifiers can identify harmful patterns in text, images, and videos. ML techniques, such as Support Vector Machines (SVM), Naïve Bayes, Random Forest, and deep learning models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have demonstrated promise in tackling the complexities of cyberbullying detection.

This paper provides a comprehensive review of various machine learning classifiers used for detecting cyberbullying. We explore their strengths, limitations, and potential improvements in accurately identifying and mitigating harmful online behavior. By understanding the efficacy of these classifiers, we aim to contribute to the development of more effective and scalable cyberbullying detection systems, ultimately promoting safer digital environments.

II. LITERATURE SURVEY

Cyberbullying detection has become a key area of research due to the increasing prevalence of online harassment. The use of machine learning classifiers to automate this process has gained considerable attention in recent years. The increasing prevalence of cyberbullying across social media platforms has made automated detection an essential area of research. With the volume of user-generated content growing exponentially, the need for efficient and accurate detection systems has become critical. Machine learning (ML) classifiers have emerged as a promising solution, enabling automated systems to detect abusive, harmful, and offensive content effectively. This section provides a detailed overview of the key contributions and approaches in the field of cyberbullying detection, focusing on various machine learning techniques employed to address this issue.

Aditya Desai et al. (2021) research pervasive use of the internet and social media has facilitated the widespread sending, receiving, and posting of negative, harmful, false, or derogatory content aimed at individuals, a phenomenon commonly known as cyberbullying. This form of bullying encompasses threats, defamation, and harassment conducted through digital platforms. The alarming rise in cyberbullying has significantly impacted mental health, particularly among the younger generation, leading to decreased self-esteem and heightened suicidal thoughts. Without effective measures to combat cyberbullying, these issues threaten to affect an entire generation of young adults. While several machine learning models have been previously employed to automatically detect cyberbullying on social media, many of these models have overlooked essential features crucial for accurately identifying or classifying bullying statements or posts. In this study, we propose a model that incorporates a comprehensive set of features essential for cyberbullying detection. Specifically, we integrate these features using a bidirectional deep learning model known as BERT.

Abdhullah-Al-Mamun et al. (2018) studied the widespread adoption of the Unicode system and the increasing use of the Internet, Bangla usage on social media platforms is on the rise. However, there has been limited research on monitoring Bangla text for social media activities due to the scarcity of annotated corpora, named dictionaries, and morphological analyzers specific to Bangla. This gap necessitates a focused exploration from the perspective of Bangladesh. This means that algorithms designed for one type of content, such as formal English, may inaccurately detect content changes, such as shifts to verbal abuse or sarcasm. Additionally, the performance of detection methods may vary due to linguistic differences between English and non-English content and the socio-emotional behaviors prevalent within the study population.To address these complexities, this paper proposes leveraging machine learning algorithms and incorporating user-specific information for detecting cyberbullying in Bangla text. A dataset of Bangla text collected from various social media platforms has been annotated, distinguishing between instances of cyberbullying and non-bullying content. This annotated dataset serves to train and evaluate different machine learning-based classification models.

Vikas S Chavan et al.(2015) studies The increasing prevalence of social networking sites among teenagers has heightened their susceptibility to cyberbullying, where computers and mobile devices are utilized for bullying activities. Comments containing abusive language can profoundly impact the psychological well-being of teens, leading to demoralization and distress. In this study, we propose methods for detecting cyberbullying using supervised learning techniques.This approach aims to improve the identification and mitigation of cyberbullying incidents on social media platforms, thereby fostering a safer online environment for teenagers.

Elif Varol Altay et al.(2018) researches on widespread use of the Internet and the accessibility of online communities, such as social media, have contributed to the rise of cybercrime. One significant manifestation of this trend is cyberbullying, a form of harassment facilitated by social networks. Cyberbullying involves sending messages containing defamatory remarks or verbally abusing individuals in front of an online audience. The unique characteristics of online social networks enable cyberbullies to reach individuals and locations that were previously inaccessible. This study employs natural language processing techniques and machine learning methods, including Bayesian logistic regression, random forest algorithm, multilayer perceptron, J48 algorithm, and support vector machines, to detect instances of cyberbullying. To the best of our knowledge, this research represents the first comprehensive comparison of these algorithms using various metrics across different experimental settings with real-world data. The aim is to enhance understanding of cyberbullying detection capabilities and contribute to the development of effective strategies for combating online harassment on social media platforms.

Mohammed Ali Al-Garadi et al (2019) critically reviews existing cyberbullying prediction models and identifies key challenges associated with their development specifically in the context of SM. It provides a comprehensive overview of the cyberbullying detection process, with detailed discussions on data collection, feature engineering, and, notably, feature selection algorithms. Various machine learning algorithms are employed to predict cyberbullying behaviors, highlighting their comparative effectiveness in different scenarios. The paper underscores the emerging issues and challenges in this field, suggesting new avenues for research exploration. By addressing these challenges, researchers can further advance the field of cyberbullying detection on social media platforms, contributing to a safer online environment for users worldwide. The misuse of social technologies, particularly on social media (SM) platforms, has introduced a new manifestation of aggression and violence that predominantly occurs in virtual spaces. This paper focuses on exploring novel forms of aggressive behavior exhibited on SM websites. It emphasizes the necessity and motivation for constructing prediction models aimed at combating such behaviors.

III. METHODOLOGY

In the detection of cyberbullying using machine learning classifiers will continue to evolve with advancements in natural language processing (NLP) and deep learning techniques. Preprocessing methods will likely become more sophisticated, utilizing cutting-edge techniques to better clean and structure text data, including contextual embeddings like BERT or GPT models, which capture nuanced language features more effectively than traditional methods like TF-IDF.

The future approach will focus on the development of more advanced feature extraction techniques, leveraging semantic and syntactic patterns to enhance the detection of subtle or disguised forms of cyberbullying. Deep learning models, such as Transformers, will likely be integrated into the pipeline, given their ability to capture long-range dependencies and contextual meanings in text, improving the classifier's ability to detect various forms of online harassment.

Additionally, future work may explore the use of ensemble methods that combine traditional machine learning algorithms with deep learning architectures to create hybrid models capable of handling diverse data sources and types. These hybrid models will be trained on more extensive and varied datasets, including data from multiple social media platforms, to improve generalization and robustness..

IV. CONCLUSION

This review highlights the growing importance of machine learning classifiers in the detection of cyberbullying, reflecting the increasing need for automated solutions to tackle online harassment. Various machine learning techniques, including decision trees, support vector machines, and deep learning models, have shown promise in identifying abusive content with significant accuracy. However, challenges such as context understanding, data imbalance, and the dynamic nature of online interactions still persist. Future research should focus on addressing these limitations by developing more context-aware models, improving datasets, and ensuring ethical fairness in detection systems. The continued evolution of these methods holds great potential for creating more robust and effective solutions to combat cyberbullying across digital platforms.

REFRENCES

[1] Aditya Desai, Shashank Kalaskar, Omkar Kumbhar and Rashmi Dhumal, "Cyber Bullying Detection on Social Media using Machine Learning" International Conference on Automation, Computing and Communication, Volume 40, 2021, DOI:

https://doi.org/10.1051/itmconf/20214003038

[2] Abdhullah-Al-Mamun and Shahin Akhter , "Social media bullying detection using machine learning on Bangla text" 2018 10th International Conference on Electrical and Computer Engineering (ICECE), DOI: 10.1109/ICECE.2018.8636797

[3] Vikas S Chavan; Shylaja S S, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network", 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), DOI: 10.1109/ICACCI.2015.7275970

[4] Elif Varol Altay; Bilal Alatas, "Detection of Cyberbullying in Social Networks Using Machine Learning Methods", 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), DOI:10.1109/IBIGDELFT.2018.8625321

[5] Mohammed Ali Al-Garadi and Mohammad Rashid Hussain, "Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges", Vol 7, IEEE, DOI: 10.1109/ACCESS.2019.2918354

[6] Rashi Shah, Srushti Aparajit, Riddhi Chopdekar, Rupali Patil, "Machine Learning based Approach for Detection of Cyberbullying Tweets", International Journal of Computer Applications, Volume 175 – No. 37, December 2020

[7] Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, Pablo García Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying", Logic Journal of the IGPL, Volume 24, Issue 1,2016,DOI: https://doi.org/10.1093/jigpal/jzv048

[8] Amgad Muneer & Suliman Mohamed Fati , "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter", Vol 12 Issue 11, MPDI,2020 , DOI: https://doi.org/10.3390/fi12110187

[9] Batoul Haidar*, Maroun Chamoun, Ahmed Serhrouchni," A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning", Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 275-284 (2017) .

[10] Manuel F. López-Vizcaíno, Francisco J. Nóvoa, Victor Carneiro, Fidel Cacheda," Early detection of cyberbullying on social media networks", Future Generation Computer Systems 118 (2021) 219–229