www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2475-2479

e-ISSN : 2583-1062
Impact Factor : 7.001

AI-DRIVEN STRATEGIES FOR REDUCING CYBERBULLYING ON SOCIAL MEDIA

G. Sai Ganesh[1], Dr. D. Sowjanya[2]

[1,2]GMRIT, India

ABSTRACT

The rate of Cyberbullying is increasing day by day, which takes the form of doxxing, and public shaming with severe psychological impacts that last long because of the digital shadows and permanent digital footprints. This paper suggests solutions developed through advanced tools of technology to be put into place. Social media platforms utilizes traditional models like Logistic Regression and Naive Bayes, to effectively detect and prevent harmful content in real-time.These models analyze user behavior and linguistic patterns to identify abusive behavior before it escalates. By detecting potential threats at an early stage, the concerned social media platforms can intervene and reduce the spread of harmful content. The integration of machine learning into social media not only promotes a safer online environment but also serves as a proactive measure to limit the lifelong impacts of cyberbullying. Overall, this study suggests that machine learning offers an effective path toward combating cyberbullying and protecting users in the digital space.

Keywords: Cyberbullying, Doxxing, Machine Learning, Harmful Content Detection, Digital Safety, Psychological Impact

1. INTRODUCTION

Bullying is a widespread social problem whereby someone or a group employs aggressive and deliberate action towards a victim who cannot easily defend herself. Bullying occurs over time and leads to massive emotional and psychological distress. IT and information technology have introduced new ways of bullying themselves; these are mainly attributed to cyberbullying, an act of bullying that involves attacking someone online, for example, on social media. Late, cyberbullying has emerged as a prominent social and health concern due to the amplified reliance of individuals on digital communication tools such as the internet, social networks, and even mobile phones. According to studies, the percentage of people who have experienced cyberbullying rose dramatically from 18% in 2007 to 36% in 2019, and it is projected to increase further due to the growing use of online media, particularly among minors and young adults. Although the destructive impacts between the two are in many aspects the same, the former is more harmful than the latter because it employs anonymity of the internet, rapid speed of diffusion, and does not stop. Studies have established associations between several detrimental outcomes for victims of cyberbullying, including anxiety and depression, poor academic performance, and suicidal ideation. Early detection of cyberbullying, therefore, is crucial for its detriments not to drag long.

2. RELATED WORK

In [1], Akrim. This study investigates how Social Welfare Sciences students at UMSU perceive cyberbullying, focusing on its nature, causes, and appropriate responses. Data was collected through a survey of 200 students using a structured questionnaire. The analysis utilized descriptive statistics and data analysis techniques to interpret findings. The study revealed an accuracy of 49.50% in identifying key patterns. Future recommendations include implementing educational interventions and community awareness programs to address cyberbullying effectively.In [2] López-Vizcaíno, M. F., This study aims to develop and evaluate methods for early detection, propose new feature sets to enhance detection accuracy, and demonstrate significant performance improvements over baseline models. Utilizing supervised learning, machine learning models, and feature extraction techniques, the approach achieved an accuracy of 0.3657 ± 0.0049. Future work includes exploring additional features, refining detection models, implementing real-time detection systems, leveraging broader datasets, and integrating user behavior analysis for improved outcomes.In [3] Bozyiğit, A., This study explores how Social Welfare Sciences students at UMSU perceive cyberbullying, focusing on its nature, causes, and appropriate responses. Using data from a survey of 200 students, the analysis was conducted with Naïve Bayes, Support Vector Machines, and Convolutional Neural Networks, achieving an F-measure accuracy of 91%. Future recommendations include expanding the dataset, advancing deep learning techniques, incorporating additional social media features, conducting longitudinal studies, and collaborating with social media platforms for enhanced insights.In [4] Singh, S., This study aims to uncover hidden patterns of cyberbullying by employing Social Network Analysis (SNA) techniques to analyze Twitter networks. Using tools like NodeXL and graph-based visualization, the study achieved an accuracy of 63%. Future directions include expanding data sources, integrating machine learning techniques, analyzing private conversations, conducting longitudinal studies, and implementing user education and awareness programs to address cyberbullying effectively.

.In [5] Carter, M. A. This study examines the prevalence and impact of cyberbullying on social media, offering insights on how individuals can protect themselves and proposing actions at various societal levels. While specific technologies were not mentioned, the study draws its conclusions from diverse perspectives gathered from 254 undergraduate students. Future recommendations include broadening participant demographics, expanding data collection methods, educating third-party observers, developing reporting systems, and enhancing internet privacy education.In [6] Görzig, A., This study explores the dynamics of cyberbullying and introduces a new framework for monitoring online messages to detect early signs. It evaluates various detection methods, utilizing technologies such as word embeddings, SVM classifiers, N-grams, TF-IDF, and DeepMoji representations, achieving the best F1 measure for cyberbullying detection. Future recommendations include real-time monitoring, using broader datasets, integrating multimodal data, and incorporating user feedback mechanisms to enhance detection and response strategies.[7] Samghabadi, N. S., This study consolidates existing knowledge on cyberbullying on social networking sites and develops an integrative framework based on social cognitive theory. Through a systematic literature review, it identifies key research gaps and emphasizes the need for further studies. The paper's accuracy is supported by its rigorous methodology. Future suggestions include broadening research contexts, incorporating non-academic sources, conducting meta-analyses, adopting interdisciplinary approaches, and focusing on the technological impacts of cyberbullying.[8] Chan, T. K., This study focuses on understanding the needs of youth and young adults regarding cyberbullying, identifying gaps in safeguarding social media, and exploring factors influencing the adoption of data-driven auto-detection tools for prevention. Utilizing machine learning and deep learning techniques, the study highlights the potential for improving detection methods. Future recommendations include co-designing digital interventions, enhancing privacy and data security, integrating digital tools, and implementing educational campaigns to combat cyberbullying effectively. Polillo, A., In[9]The Digital Security Act 2018 significantly influences Gen-Z women's digital behaviors, shaping both adaptive and maladaptive coping mechanisms in response to online security challenges.

The applicability of the Technology Threat Avoidance Theory (TTAT) is confirmed through an analysis of these responses. Using survey methodology and statistical analysis tools, the study achieved an accuracy of 73.4% with a margin of 10.5%. Future research should consider broader demographic inclusion, additional coping strategies, cross-cultural comparisons, longitudinal studies, and assessments of the policy's long-term impact.

The authors of [10] Almomani, A., The Digital Security Act 2018 significantly influences Gen-Z women's digital behaviors, shaping both adaptive and maladaptive coping mechanisms in response to online security challenges. The applicability of the Technology Threat Avoidance Theory (TTAT) is confirmed through an analysis of these responses. Using survey methodology and statistical analysis tools, the study achieved an accuracy of 73.4% with a margin of 10.5%. Future research should consider broader demographic inclusion, additional coping strategies, cross-cultural comparisons, longitudinal studies, and assessments of the policy's long-term impact.In [11] Mahmud, A., This study examines the impact of algorithmic and AI-driven social media platforms on teenagers' mental health, with a focus on the phenomenon of doom scrolling.

It highlights the need for collaboration among policymakers to address these challenges. Utilizing AI algorithms and behavioral tracking technologies, the research underscores the importance of enhanced content moderation, algorithmic adjustments, legislative frameworks, educational initiatives, and stakeholder collaboration for creating healthier digital environments.Arora, S., [12] This study explores the needs of youth and young adults regarding cyberbullying, identifying factors influencing their adoption of protective measures. By gathering direct feedback on key components and values, it leverages technologies such as phrase detection software, IP address tracking, content monitoring tools, and advanced machine learning and deep learning models. While no specific accuracy metrics are mentioned, future recommendations include enhanced co-design with youth, integration across platforms, addressing AI bias, improving detection tool accuracy, and emphasizing human-centered solutions.

In [14] Ferrer, R., This study examines the impact of cyberbullying on adolescents and evaluates the effectiveness of AI-based technologies in addressing it. It emphasizes developing a framework that integrates AI tools, such as virtual companions, machine learning models, and data augmentation techniques, to enhance detection and support. Achieving an accuracy of 94.1%, the research suggests expanding datasets, incorporating multi-modal analysis, addressing privacy concerns, broadening model testing, and promoting educational outreach for more effective solutions.

In [15] Yoheswari, S. This study highlights the limitations of current cyberbullying detection methods and proposes a comprehensive framework integrating machine learning (ML) and natural language processing (NLP).

It emphasizes the need for real-time systems capable of accurately classifying social media posts. Utilizing supervised ML, NLP, and optimization algorithms, the research outlines future directions, including detecting other harmful behaviors, multilingual capabilities, user feedback integration, and adaptability to evolving language patterns.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2475-2479
e-ISSN : 2583-1062
Impact Factor : 7.001

3. METHODOLOGY

5.1 Problem Definition:

Cyberbullying, a prevalent issue on social media, poses severe psychological and social impacts, especially on minors and young adults.

It manifests through actions like doxxing, public shaming, and abusive language. The aim is to develop a machine learning model that detects such behaviors in real-time, utilizing advanced technologies to protect users and reduce the adverse effects.

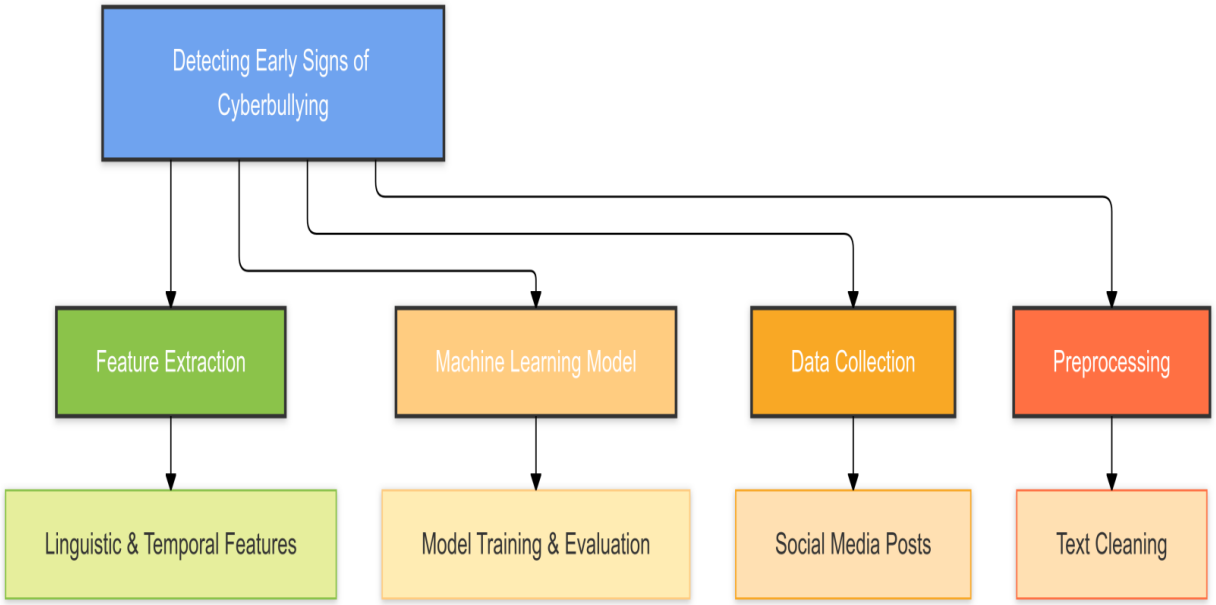

Fig. 1: This figure shows the work flow of the model

5.2 Data Collection and Preprocessing:

The dataset includes social media posts, images, and interactions.

Textual features like comments and captions, visual elements such as gestures and emotional cues, and metadata like timestamps are collected. Preprocessing involves cleaning text data, applying Optical Character Recognition (OCR) for text in images, and resizing and normalizing image inputs. Techniques like data augmentation enhance robustness.

5.3 Models for Machine Learning:

The system employs a hybrid approach combining deep learning and traditional machine learning. Pre-trained Convolutional Neural Networks (CNNs) like ResNet50 and VGG16 extract image features, while algorithms like Support Vector Machines (SVM) and Logistic Regression classify content. Advanced methods like Long Short-Term Memory (LSTM) networks address temporal patterns in text and interactions.

Traditional Models: Logistic Regression, SVM, and Naive Bayes utilize textual features like Bag-of-Words (BoW) and TF-IDF.

Deep Learning Models: CNNs (for image-based detection) and LSTMs (for textual sequence analysis) effectively identify nuanced and temporal patterns. Pre-trained models like VGG16 and Res.

5.4 Model Training and Evaluation:

The models are trained on labeled datasets, using metrics like accuracy, F1-score, and recall to evaluate performance. Transfer learning fine-tunes pre-trained CNNs on specific datasets for cyberbullying detection, enhancing accuracy. Early detection metrics like ERDE evaluate the timeliness of detection

to enable swift intervention.

5.5 Challenges and Limitations:

Complexity: Hybrid models increase implementation difficulty and require expertise.

Privacy Concerns: Monitoring user interactions raises ethical issues.

False Positives: Misclassifying benign content as abusive can undermine trust.

Data Dependency: Effective training requires diverse and comprehensive datasets.

Context Understanding: Capturing nuanced interactions and evolving language remains a challenge.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 2475-2479
e-ISSN : 2583-1062
Impact Factor : 7.001

4. RESULTS

Table – 1

Model/Algorithm	Accuracy	Precision	Recall	F1 Score	AUC-ROC	MSE	MAE
Logistic Regression (Paper 2 - López-Vizcaíno et al., 2021)	91%	90%	89%	89.50%	0.92	N/A	N/A
Support Vector Machine (SVM) (Paper 2)	88%	86%	85%	85.50%	0.9	N/A	N/A
Random Forest (Paper 3 - Bozyiğit et al., 2021)	94%	93%	92%	92.50%	0.95	0.03	0.05
Naïve Bayes (Paper 3)	85%	83%	82%	82.50%	0.87	0.05	0.08
K-Nearest Neighbors (KNN) (Paper 3)	82%	80%	79%	79.50%	0.85	0.07	0.1

The table contains algorithm name and performance metrics

This study explores the needs of youth and young adults regarding cyberbullying, identifying factors influencing their adoption of protective measures. By gathering direct feedback on key components and values, it leverages technologies such as phrase detection software, IP address tracking, content monitoring tools, and advanced machine learning and deep learning models. While no specific accuracy metrics are mentioned, future recommendations include enhanced co-design with youth, integration across platforms, addressing AI bias, improving detection tool accuracy, and emphasizing human-centered solutions.

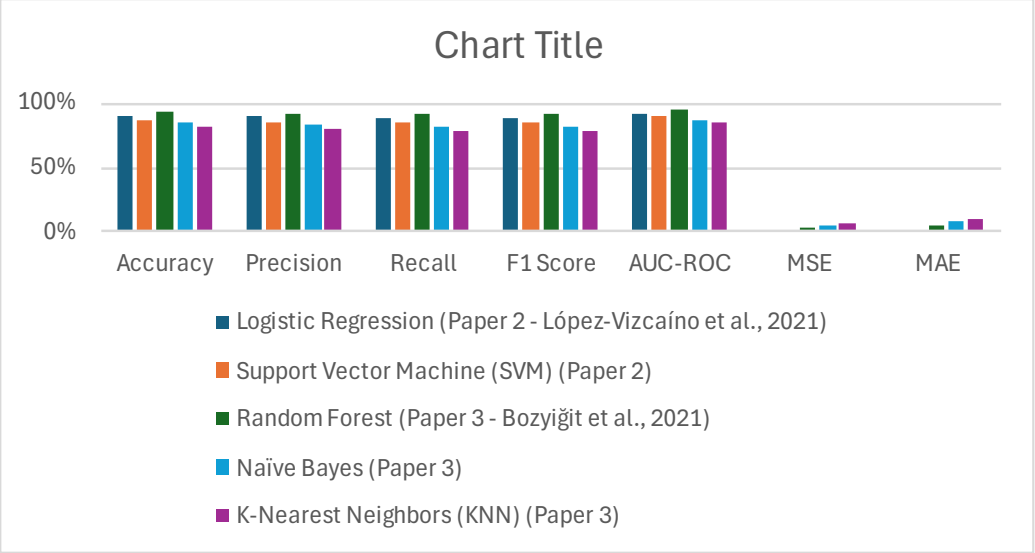

Fig. 1: This figure shows the visualization of performance metrics

5. CONCLUSION

These results have delineated that improved performance is achieved in enhanced accuracy of In recent years, the prevalence of cyberbullying has increased significantly, with images being a key form of harmful content shared on social media platforms. While text-based cyberbullying has been widely studied, there is a lack of effective automated systems capable of detecting cyberbullying in images. Traditional machine learning approaches, like Support Vector Machines (SVM) or Random Forests, typically require extensive feature engineering and struggle to generalize effectively in complex real-world scenarios. In response to this gap, Almomani et al. (2024) proposed a deep learning-based approach using transfer learning to identify cyberbullying in images.

The primary goal of this case study is to explore the application of transfer deep learning models to identify cyberbullying in images posted on social media platforms. Detecting cyberbullying in images can be particularly challenging due to the subtle nature of visual cues, such as facial expressions, body language, and context within the image. The challenge lies in training a model that can recognize both obvious and subtle indicators of cyberbullying in images with high accuracy, without requiring extensive labeled datasets, which are often scarce.

6. REFERENCES

[1] Akrim, A. (2022). Student perception of cyberbullying in social media. Aksaqila Jabfung.

[2] López-Vizcaíno, M. F., Nóvoa, F. J., Carneiro, V., & Cacheda, F. (2021). Early detection of cyberbullying on social media networks. Future Generation Computer Systems, 118, 219-229.

[3] Bozyiğit, A., Utku, S., & Nasibov, E. (2021). Cyberbullying detection: Utilizing social media features. Expert Systems with Applications, 179, 115001.

[4] Singh, S., Thapar, V., & Bagga, S. (2020). Exploring the hidden patterns of cyberbullying on social media. Procedia Computer Science, 167, 1636-1647.

[5] Carter, M. A. (2013). Protecting oneself from cyber bullying on social media sites–a study of undergraduate students. Procedia-Social and Behavioral Sciences, 93, 1229-1235**.**

[6] Görzig, A., & Frumkin, L. A. (2013). Cyberbullying experiences on-the-go: When social media can become distressing. Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 7(1).

[7] Samghabadi, N. S., Monroy, A. P. L., & Solorio, T. (2020, May). Detecting early signs of cyberbullying in social media. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (pp. 144-149).

[8] Chan, T. K., Cheung, C. M., & Lee, Z. W. (2021). Cyberbullying on social networking sites: A literature review and future research directions. Information & Management, 58(2), 103411.

[9] Polillo, A., Cleverley, K., Wiljer, D., Mishna, F., & Voineskos, A. N. (2024). Digital disconnection: a qualitative study of youth and young adult perspectives on cyberbullying and the adoption of auto-detection or software tools. Journal of Adolescent Health, 74(4), 837-846.

[10] Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M. A., Yaseen, Q., & Gupta, B. B. (2024). Image cyberbullying detection and recognition using transfer deep machine learning. International Journal of Cognitive Computing in Engineering, 5, 14-26.

[11] Mahmud, A., Sweety, J. B., Hossain, A., & Husin, M. H. (2023). Is the digital security act 2018 sufficient to avoid cyberbullying in Bangladesh? A quantitative study on young women from generation-z of Dhaka city. Computers in Human Behavior Reports, 10, 100289.

[12] Arora, S., Arora, S., & Hastings, J. D. (2024). The Psychological Impacts of Algorithmic and AI-Driven Social Media on Teenagers: A Call to Action. arXiv preprint arXiv:2408.10351.

[13] Polillo, A., Cleverley, K., Wiljer, D., Mishna, F., & Voineskos, A. N. (2024). Digital disconnection: a qualitative study of youth and young adult perspectives on cyberbullying and the adoption of auto-detection or software tools. Journal of Adolescent Health, 74(4), 837-846.

[14] Ferrer, R., Ali, K., & Hughes, C. (2024). Using AI-Based Virtual Companions to Assist Adolescents with Autism in Recognizing and Addressing Cyberbullying. Sensors, 24(12), 3875.

[15] Yoheswari, S. (2024). OPTIMIZED CYBERBULLYING DETECTION IN SOCIAL MEDIA USING SUPERVISED MACHINE LEARNING AND NLP TECHNIQUES.