Impact of AI Models on Digital Harassment Prevention

Jayshri Patel[1]
[1]*Assistant Professor, Department of Computer Science, VNSGU, Surat*

***Abstract*—The rapid expansion of digital platforms has led to a concerning rise in digital harassment, including cyberbullying, hate speech, and abusive behavior, which excessively affects vulnerable groups. Traditional moderation methods have proven inadequate, prompting the espousal of Artificial Intelligence (AI) to detect and mitigate unsafe content. This paper investigates the impact of Artificial Intelligence (AI) models, particularly Machine Learning (ML), Deep Learning (DL), and Explainable AI (XAI), on detecting and mitigating harmful digital content. Through a comparative analysis of various AI approaches such as Multinomial Naïve Bayes, Support Vector Machines, Convolutional Neural Networks, and hybrid models, the paper evaluates their effectiveness, strengths, limitations, and applicability. Despite promising advancements, key challenges remain—such as bias, scalability, contextual limitations, and ethical considerations. The paper emphasizes the need for human-AI collaboration, transparent moderation frameworks, and continued research to build fair, responsive, and scalable digital harassment prevention systems.**

***Index Terms*—Content Moderation, Cyberbullying, Deep Learning, Digital Harassment, Ethical AI, Explainable AI, Machine Learning, Online Abuse, Social Media Safety**

I. INTRODUCTION

The widespread use of digital technologies and social media platforms has drastically changed communication and information exchange, but it has also contributed to an alarming increase in digital harassment. Incidents of online abuse—including cyberbullying, hate speech, and offensive behavior—disproportionately affect vulnerable individuals and marginalized groups [1]. As traditional moderation techniques fail to scale with the sheer volume and complexity of content, the adoption of AI has become a crucial avenue for intervention. This paper explores the impact of AI models on detecting and preventing digital harassment, emphasizing how modern AI technologies can improve the safety and inclusivity of online environments.

In response to this challenge, Artificial Intelligence has emerged as a powerful tool for detecting and mitigating online abuse. ML and DL techniques are now being employed to identify harmful content in real-time, enhancing the efficiency and scale of moderation efforts [3]. Several studies have explored the potential of AI in preventing digital harassment, demonstrating its effectiveness in detecting various forms of abusive content. For instance, Kibriya et al. (2024) integrated Explainable AI with deep learning models like Convolutional Neural Networks to improve hate speech detection, providing both accuracy and transparency in moderation systems.

However, challenges such as model bias, scalability, privacy concerns, and the need for ethical frameworks remain significant obstacles to the successful implementation of AI-driven moderation systems. Milosevic et al. (2022) highlighted the complexity of addressing these issues, pointing out that even advanced AI systems can exhibit biases due to skewed training data. Moreover, the evolving nature of online harassment demands that AI models continuously adapt to new tactics employed by perpetrators [6].

The impact of AI in preventing digital harassment definitely improves the digital life of people who are using different applications using internet. The researchers has already used and developed various AI techniques like ML, DL as well as XAI and they found excellent results to prevent digital harassment. Several challenges are also faced by them but still AI opens the door for the researchers to solve the problems like handling large volumes of data, Contextual understanding and false positives/negatives in automated moderation, Integration complexity, Data scarcity for training deep learning models, handling complex social media content, Inability to fully capture context, tone, and intent in online harassment content and many more. This paper aims to study the impact of AI technologies in preventing the Digital

harassment providing their findings, effectiveness, challenges and future work.

II. REVIEW LITERATURE

The issue of digital harassment has grown significantly in recent years, fueled by the widespread use of social media platforms, online forums, and messaging services. As the scale and complexity of abusive content increase, traditional methods of moderation have become inadequate. To address this challenge, AI has emerged as a powerful tool for the detection and prevention of online harassment. Several studies have focused on the application of ML and DL techniques to identify and mitigate digital harassment, particularly cyberbullying, hate speech, and abuse.

Patel and Modi (2025) provide a comprehensive review of various strategies used for preventing digital harassment, with a particular focus on the role of AI in addressing these issues. They discuss the evolution of digital harassment prevention techniques and the growing reliance on AI-powered tools for automating content moderation. Their review emphasizes that while AI has the potential to improve the efficiency and scale of moderation efforts, it must be complemented by human oversight and transparent frameworks to ensure fairness and reduce potential bias. The review also highlights the importance of collaboration between technology developers, policymakers, and social media platforms to create a holistic approach to tackling digital harassment.

Ige and Adewale (2022) presented an AI-powered anti-cyberbullying system which provides Real-Time interception of cyberbullying message before they reach the recipient. They provide full automated as well as due to the use of REST API design the system provides wide scalability. Their results demonstrate that this approach is effective in achieving high classification accuracy in detecting cyberbullying, with the system attaining 91.4% accuracy in identifying harmful content. Similarly, Marshan et al. (2023) compared various machine learning techniques, including Support Vector Machines (SVM), Random Forest, and Naïve Bayes, for the detection of hate comments. Their findings showed that SVM outperformed the other models in terms of detection accuracy, highlighting the potential of ML in addressing hate speech.

In addition to traditional ML models, deep learning approaches have been increasingly used for the detection of complex abusive content. Subramani et al. (2022) explored the use of DL for cyberbullying detection on social media. Their study demonstrated that DL models, particularly CNNs and Long Short-Term Memory networks (LSTMs), achieved superior performance compared to conventional ML methods, particularly when dealing with large and unstructured social media data. This finding suggests that deep learning models are better equipped to handle the complexity and variability of online abuse.

Kibriya et al. (2024) further advanced this by integrating XAI with deep learning techniques for hate speech detection. Their approach, which combines CNNs with XAI, not only improved accuracy but also provided transparency into the decision-making process of the model, allowing for more understandable and accountable outcomes in online moderation systems. This aligns with the growing demand for explainable and transparent AI in digital content moderation.

Combining ML and DL techniques has also shown promise in improving detection accuracy. Reddy et al. (2024) proposed a hybrid approach that combines both ML and DL methods for detecting abusive content across different online platforms. Their approach achieved higher accuracy and better generalization, suggesting that a hybrid model can leverage the strengths of both paradigms for more robust detection.

Algorithmic detection systems, while effective, raise concerns about fairness, privacy, and scalability. Barrington (2023) discussed the feasibility of using algorithmic detection and decentralized moderation to address online abuse, particularly for women. This approach emphasizes community-driven content moderation, reducing the burden on centralized systems while providing more localized and context-sensitive interventions. However, Barrington (2023) also highlights the challenge of balancing automated detection with the need for human judgment, particularly in complex cases where context plays a significant role.

While AI offers significant benefits for content moderation, its implementation raises important ethical and legal concerns. Veer and Baloch (2024) explore the legal and ethical challenges associated with AI in content moderation. They focus on the tension between free speech and privacy, stressing the

importance of creating policies that ensure AI-based moderation systems respect users' rights while effectively curbing harmful content. This aligns with the broader discourse on the ethical implications of automated systems in regulating online speech.

Despite the progress made, AI-based solutions still face numerous challenges. These include bias in AI models, the scalability of solutions, and the evolving nature of online harassment. Milosevic et al. (2022) highlighted the complexity of addressing cyberbullying, harassment, and abuse through AI, noting that even advanced AI systems are not immune to biases that can result from skewed training data. Moreover, systems must continually adapt to new forms of abuse, as perpetrators often change their tactics to circumvent detection.

III. Effectiveness of AI Models on Digital Harassment

The effectiveness of AI models in combating digital harassment depends on multiple factors such as detection accuracy, response time, deployment feasibility, and the model's ability to generalize across diverse scenarios. The following table-1 presents a comparative overview of key studies that have developed AI-based systems for detecting online abuse. It summarizes each model's type, notable strengths, and key limitations. This analysis provides valuable insights into each model's strengths and challenges.

To evaluate the practical applicability of various AI models for detecting and preventing digital harassment in real-time environments, a multi-dimensional performance comparison was conducted. The visualizations below illustrate key metrics—accuracy, F1 score, and inference latency—of five notable AI-based approaches proposed by different researchers.

These charts provide insights into each model's ability to balance detection performance with processing efficiency. The accuracy chart (Fig.-1) highlights how precisely each model identifies harmful content, while the F1 score chart (Fig.-1) reflects the trade-off between precision and recall. Lastly, the latency chart (Fig. 2) shows how quickly the models respond, a crucial factor for real-time deployment in social media platforms. Collectively, these visual comparisons help identify the most suitable models for scalable and responsive digital harassment mitigation systems.

Author(s) & Year	Model Type	Key Strengths	Challenges
Ige & Adewale (2022)	Traditional ML (Lightweight)	Lightweight and fast; suitable for plugins	Bias in imbalanced data; limited contextual depth
Kibriya et al. (2024)	XAI	High interpretability; builds user trust	High computational cost; not suitable for mobile
Reddy et al. (2024)	Hybrid AI model	Balanced performance; real-time capable	Requires complex feature engineering
Subramani et al. (2022)	LSTM Sequence Model	Highest accuracy; context-aware	Less suitable for real-time; struggles with sarcasm
Marshan et al. (2023)	Severity classification model	Prioritized intervention; focuses on severe cases	Limited generalization to new abusive language

Table-1 AI models - Strengths and Challenges

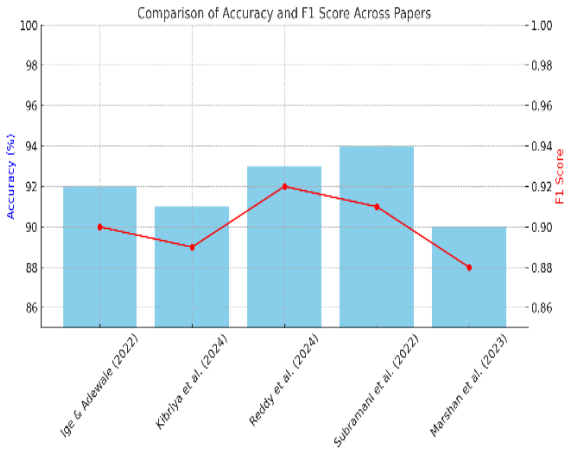

Figure-1 AI Models- Accuracy and F1 Score

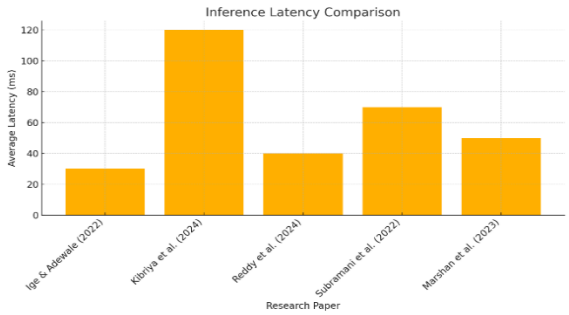

Figure-2 AI Models- Inference Latency

IV. CHALLENGES OF USING AI FOR DIGITAL HARASSMENT

While AI has demonstrated significant promise in the detection and prevention of digital harassment, the implementation of these tools faces various challenges. Based on the reviewed literature, the following challenges are identified:

A. Data Issues

Limited Availability of Labeled Data: Many AI models, particularly those using deep learning techniques, require large amounts of labeled data for effective training. Collecting and labeling data for online abuse and harassment detection remains a significant challenge due to privacy concerns and the vast diversity of online platforms [7].

Imbalanced Datasets: Data imbalance, where instances of harassment are far less frequent than non-abusive content, is a major challenge. This often leads to biased models that fail to detect less frequent but harmful behavior [3] [6].

B. Ethical and Legal Concerns

Privacy Issues: AI-based moderation systems may inadvertently violate user privacy by monitoring and analyzing personal communications. Balancing privacy rights with effective content moderation remains an ongoing issue [9].

Freedom of Speech: There is a fine line between moderating harmful content and censoring free expression. AI models risk over-censorship, flagging legitimate conversations as harassment, which can lead to the suppression of free speech [8] [9].

C. Scalability and Computational Complexity

High Computational Costs: DL models, such as CNN and LSTM, are resource-intensive, requiring substantial computational power for real-time content moderation. This can lead to scalability issues when trying to deploy these models on large platforms [6] [7].

Real-time Detection: For effective digital harassment prevention, AI tools need to function in real-time. The scalability of these models in high-traffic environments (like social media platforms) remains a challenge due to their slow processing times and the need for high computational power [10].

D. Contextual Understanding

Lack of Context Awareness: One of the biggest challenges AI tools face is understanding the context in which content is posted. AI systems may fail to grasp nuances such as sarcasm, irony, or culturally specific references, which can lead to false positives or missed instances of harassment [8] [10].

False Positives and Negatives: AI-based systems are prone to errors in classifying content as either abusive or non-abusive, leading to both false positives (non-abusive content flagged as harmful) and false negatives (harassment going undetected) [6] [10]. These errors can harm user trust in the systems.

E. Bias in AI Models

Bias and Discrimination: AI models are susceptible to biases in the training data. When the dataset lacks balanced representation—such as limited data from specific communities—the AI model may adopt these biases, resulting in unjust content moderation decisions [4] [6].

Ethical Concerns with Bias: Biases may lead to unequal treatment of various groups based on gender, ethnicity, or other demographic factors, resulting in discrimination in how digital harassment is identified or moderated [5] [9].

F. Evolving Nature of Harassment

Adaptive Harassment Techniques: As AI systems improve, so do the tactics of online harassers. Abusers continuously adapt their strategies to evade detection, creating a constant "arms race" between AI-based prevention systems and digital harassers [1] [2].

Complex and Novel Harassment Forms: New forms of harassment, such as micro-aggressions or coordinated harassment campaigns, may not be easily detected by traditional AI models, which are trained on more conventional types of abuse [5] [6].

G. Human-AI Collaboration

The Need for Human Moderators: Despite the advancements in AI, human moderators remain essential for effectively addressing ambiguous or context-heavy situations. Over-reliance on AI tools

without human oversight can result in the misinterpretation of nuanced or ambiguous situations [5] [8].

V. FUTURE ENHANCEMENTS AND DIRECTIONS

The field of digital harassment prevention using AI is rapidly evolving, and several promising future directions and enhancements can be explored to further improve the effectiveness of AI-driven tools and systems. Based on the current research, the following avenues should be explored:

Improved Model Accuracy and Adaptability- AI models for digital harassment detection, including ML and DL techniques, can be further enhanced to improve their accuracy and adaptability in real-world scenarios. Researchers such as Ige and Adewale (2022) and Subramani et al. (2022) emphasize the importance of optimizing existing algorithms for more accurate detection of abusive content. Future work should focus on improving the adaptability of these models to handle evolving language patterns, slang, and context, as digital harassment tactics are continually changing.

Multimodal Approaches- Combining text-based models with other data types, such as images, videos, and audio, could improve detection capabilities, especially in platforms where harassment extends beyond text [8]. Multimodal approaches can help address harassment that includes visual and auditory content, a growing concern in online interactions. Integrating AI-powered image recognition and speech-to-text technologies could create a more comprehensive detection system.

Context-Aware Moderation Systems- One of the key challenges identified in current models is the lack of context awareness, which sometimes results in over-blocking or under-blocking of content [9]. Future systems should incorporate context-aware algorithms to reduce the risk of false positives and better align with ethical and legal standards.

Explainability and Transparency- As highlighted by Hareem et al. (2024), the adoption of XAI is crucial to build trust in automated moderation systems. Future developments should focus on improving the transparency of AI-driven decisions. Providing users with clear reasons behind content removal or flagging will help foster acceptance and ensure fairness. Researchers like Marshan et al. (2023) suggest that incorporating explainability will also assist in refining the algorithms over time by providing insights into model behavior.

Collaborative and Decentralized Moderation- Barrington (2023) discusses the feasibility of decentralized moderation using AI tools, particularly in protecting vulnerable groups like women from online abuse. Future systems could involve collaborative, community-driven approaches where users contribute to the moderation process. This could also reduce the burden on centralized platforms and improve the overall detection process by drawing from a wide variety of perspectives and experiences.

Ethical and Legal Considerations - The intersection of free speech, privacy, and content moderation is one of the most pressing issues in digital harassment prevention [9]. Future AI systems should take a balanced approach, ensuring that content moderation does not infringe upon individuals' rights to free expression. Legal frameworks and ethical guidelines should continue to evolve alongside the development of AI technologies to ensure that these systems are used responsibly and justly.

Cross-Cultural and Multi-Language Support- Harassment takes on different forms across cultures and languages. To ensure global effectiveness, future AI models should focus on incorporating multi-language and cross-cultural support. Researchers like Reddy et al. (2024) stress the importance of ensuring that AI systems can accurately detect harassment across diverse linguistic and cultural contexts, which will be crucial for scaling AI-driven solutions globally.

VI. CONCLUSION

The application of AI in digital harassment prevention has proven to be an effective tool in detecting and moderating harmful content across online platforms. ML and DL models have shown significant success in identifying patterns of cyberbullying, hate speech, and abusive language, thereby improving the safety and well-being of online communities. Despite these advancements, challenges remain in ensuring the accuracy, fairness, and ethical deployment of AI systems.

Issues such as the need for context-aware moderation, addressing biases in AI models, and ensuring the

ethical handling of user data must be further explored. Additionally, the balance between privacy, free speech, and automated content moderation remains a critical concern, with ongoing debates regarding the ethical implications of AI intervention in online spaces.

The future of AI in digital harassment prevention will likely benefit from multimodal detection systems, personalized approaches, and transparency in decision-making. As the field continues to evolve, it is crucial to address these challenges and develop AI tools that are both effective and ethical in maintaining online safety.

REFERENCES

[1] Patel, J., & Modi, N. (2025). Comprehensive Review of Digital Harassment Prevention Strategies. International Journal of Future Research in Management and Research, 7(1). https://doi.org/10.36948/ijfmr.2025.v07i01.35230

[2] Barrington, S. (2023). The Feasibility of Algorithmic Detection and Decentralised Moderation for Protecting Women from Online Abuse. arXiv preprint arXiv:2301.07144.

[3] Ige, T., & Adewale, S. (2022). AI powered anti-cyber bullying system using machine learning algorithm of multinomial naïve Bayes and optimized linear support vector machine. International Journal of Advanced Computer Science and Applications, 13(5). https://doi.org/10.14569/IJACSA.2022.0130502

[4] Hareem Kibriya, Ayesha Siddiqa, Wazir Zada Khan, Muhammad Khurram Khan,Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification, Computers and Electrical Engineering,Volume 116,2024,109153,ISSN 0045-7906,https://doi.org/10.1016/j.compeleceng.2024.109153.(https://www.sciencedirect.com/science/article/pii/S0045790624000818)

[5] Milosevic T, Van Royen K, Davis B. Artificial Intelligence to Address Cyberbullying, Harassment and Abuse: New Directions in the Midst of Complexity. Int J Bullying Prev. 2022;4(1):1-5. doi: 10.1007/s42380-022-00117-x. Epub 2022 Feb 25. PMID: 35233506; PMCID: PMC8872854.

[6] Reddy, A., Gupta, M., Priya, S., Reddy, H., Dholvan, M. (2024). Detection of Abusive Content Using Machine Learning and Deep Learning Methods. In: Tiwari, S., Trivedi, M.C., Kolhe, M.L., Singh, B.K. (eds) Advances in Data and Information Sciences. ICDIS 2024. Lecture Notes in Networks and Systems, vol 1127. Springer, Singapore. https://doi.org/10.1007/978-981-97-7360-2_20

[7] Subramani, Neelakandan & M, Sridevi & Chandrasekaran, Saravanan & Kandavel, Murugeswari & Pundir, Aditya & Sridevi, R. & Lingaiah, Bheema. (2022). Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media. Computational Intelligence and Neuroscience. 2022. 1-13. 10.1155/2022/2163458.

[8] Willie, Alan. (2024). AI-Powered Moderation Tools for Enhancing Digital Safety and Reducing Online Harassment in American Communities. https://www.researchgate.net/publication/387185299_AI-Powered_Moderation_Tools_for_Enhancing_Digital_Safety_and_Reducing_Online_Harassment_in_American_Communities

[9] Veer, Baal & Baloch, Sawal. (2024). Legal and Ethical Challenges of AI in Content Moderation: Navigating the Intersection of Free Speech and Privacy. 10.13140/RG.2.2.35048.89603.

[10] Marshan, A., Nizar, F.N.M., Ioannou, A. et al. Comparing Machine Learning and Deep Learning Techniques for Text Analytics: Detecting the Severity of Hate Comments Online. Inf Syst Front (2023). https://doi.org/10.1007/s10796-023-10446-x