Virtual Guard: AI-Driven Defense Against Harmful Digital Content

Ilma Naaz, Department of Computer Applications, Invertis University, Bareilly, India, ilmanaaz3006@gmail.com
Yusra Khan, Department of Computer Applications, Invertis University, Bareilly, India, yusrakhan54545@gmail.com
Atul Kumar, Department of Computer Applications, Invertis University, Bareilly, India, atulkr@gmail.com
Navnika Kapoor, Department of Computer Applications, Invertis University, Bareilly, India, navnikakapoor344@gmail.com
Deepali Mishra, Department of Computer Applications, Invertis University, Bareilly, India, deep.mis25@gmail.com

Abstract- Artificial Intelligence (AI)-based content moderation is widely used in social media platforms to filter out harmful and inappropriate content, including hate speech, misinformation, and explicit material. This paper explores existing AI moderation techniques, their effectiveness, and current loopholes. The study highlights challenges such as adversarial attacks, bias in AI models, scalability, and privacy concerns. Finally, we propose solutions including adversarial training, cross-lingual transformers, explainable AI (XAI), and federated learning for privacy-preserving moderation. The exponential growth of social media platforms has resulted in a surge of user-generated content, necessitating robust moderation mechanisms to filter harmful and sensitive material. Manual moderation is no longer scalable due to psychological toll and inefficiency. This paper explores the role of Artificial Intelligence (AI) in content moderation, focusing on the types of sensitive content, their impact on victims, existing loopholes in current systems, and proposes novel solutions to enhance moderation accuracy. We also reference real-life incidents to validate the need for stronger AI moderation and outline areas still under-researched.

Keywords- social media, moderation, sensitive content, artificial intelligence crimes, victims, adversarial attacks.

1. Introduction

Social media has become a primary source of communication, but it also facilitates the spread of harmful content. AI-based content moderation uses machine learning techniques to automate the filtering of inappropriate material. [1], [2] However, current moderation systems face limitations in understanding context, handling adversarial attacks, and ensuring fairness. This paper aims to analyze existing AI-based moderation methods, identify their shortcomings, and propose solutions to enhance their effectiveness.

With billions of users online, social media platforms like Facebook, Instagram, YouTube, and Twitter host an overwhelming amount of content daily. This content ranges from positive engagement to harmful and offensive materials such as hate speech, sexual abuse, and fake news. Manual moderation cannot keep pace, leading to the integration of AI-based systems for efficient and real-time moderation. [3]

2. Research Objectives

- To analyse existing AI-based content moderation techniques.
- To identify current loopholes in these techniques and assess their impacts.
- To propose effective solutions that can address and resolve these challenges.

3. Literature Review- Existing AI-based content moderation techniques include:

- Natural Language Processing (NLP): Models like BERT, GPT, and T5 detect hate speech and abusive content.
- Computer Vision: CNNs (Convolutional Neural Network) and vision transformers identify explicit images and violent content.
- Hash-Matching: Comparing content with a database of known harmful material.
- User Behavior Analysis: Detecting repeated violations and suspicious posting patterns.

4. Types of Sensitive Content and Real-World Impact

According to the Pew Research Center (2023) as shown in figure 2, over 60% of teenagers in the U.S. have experienced some form of cyberbullying. [4] as describe in figure 1, In India, a 2022 NCRB report highlighted over 11,000 cases linked to cyber harassment. [5]

4.1. Exposure to Pornographic & Sexually Explicit Content

4.1.1 Impact:

- Early exposure increases sexual risk-taking, mental health issues, and unrealistic body image standards, especially among youth. [6]

4.1.2 Real Incidents:

- UK Study (NSPCC 2022): 88% of teens had seen sexual content online by age 14. [7]
- OnlyFans & Twitter (2022): Rise in explicit content on platforms led to concerns about underage exposure as shown in figure 1. [8]

4.1.3 Consequences:

- Anxiety, depression, sexual aggression, body image issues.
- Decreased academic performance in teens (CDC study). [9]

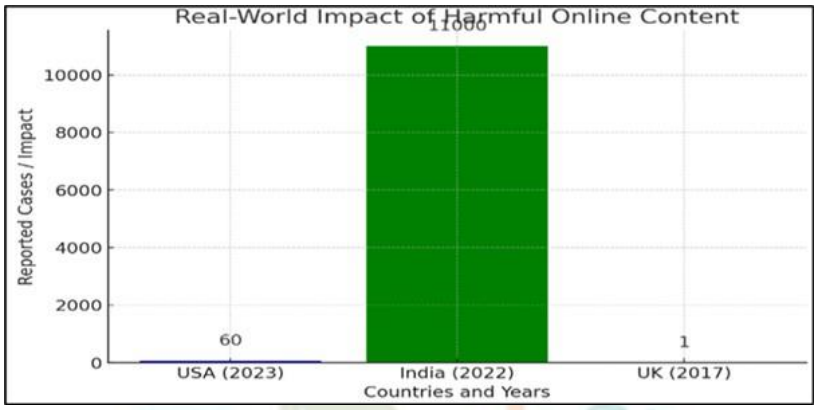

Figure 1: Figure for showing the numeric graphics of World Impact of harmful content on social media.

4.2. Hate Speech & Racism

4.2.1 Impact:

- Fuels discrimination, violence, and social divide.

4.2.2 Real Incidents:

- Rohingya Genocide (Myanmar, 2016–17): Facebook was used to incite hatred against Rohingya Muslims. UN accused Facebook of enabling violence.
- Black Lives Matter (2020): Surge in racist trolling, misinformation, and targeted attacks.

4.2.3Consequences:

- Real-life mob lynchings, communal violence, and increased mental trauma in targeted communities.

4.3. Cyberbullying

4.3.1 Impact:

- Major contributor to depression, self-harm, and suicide among youth.

4.3.2 Real Incidents:

- Megan Meier Case (2006): 13-year-old died by suicide after online bullying on MySpace.
- Amanda Todd (2012, Canada): Posted a video about her bullying experience, died by suicide.

4.3.3 Stats:

- CDC (2023): 1 in 5 teens in the US reported cyberbullying.
- UNESCO: Girls are 27x more likely to be harassed online.

4.4 Suicide & Self-Harm Encouraging Content

4.4.1 Impact:

- Normalizes or glorifies self-harm, pushing vulnerable people to attempt suicide.

4.4.2 Real Incidents:

- TikTok “Blue Whale Challenge” & “Momo Challenge”: Dangerous trends that encouraged children to harm themselves.
- Molly Russell Case (UK, 2017): 14-year-old took her life after consuming self-harm content on Instagram. Meta faced global scrutiny.

4.4.3 Stats:

- WHO (2023): Suicide is the 4th leading cause of death in 15–29 age group.

4.5. Violent Content & Extremism

4.5.1 Impact:

- Radicalizes viewers, spreads fear, incites copycat crimes.

4.5.2 Real Incidents:

- Christchurch Mosque Attack (New Zealand, 2019): Livestreamed on Facebook; took 17 minutes to remove.
- Capitol Riots (USA, 2021): Platforms used to organize and incite political violence.

4.5.3 Consequences:

- Platforms were blamed for lack of proactive moderation.
- Massive policy reforms in Australia, EU, and US.

4.6. Misinformation & Fake News

4.6.1 Impact:

- Affects public opinion, elections, vaccine resistance, and communal harmony as shown in figure 2.

4.6.2 Real Incidents:

- COVID-19 Pandemic: Facebook & WhatsApp were flooded with fake cures and anti-vaccine content.
- Indian WhatsApp Lynchings (2018): Fake child abduction messages led to mob killings.

4.6.3 Stats:

- WHO termed misinformation an "infodemic."
- Over 800 deaths globally in early 2020 due to fake COVID remedies (Johns Hopkins Report).

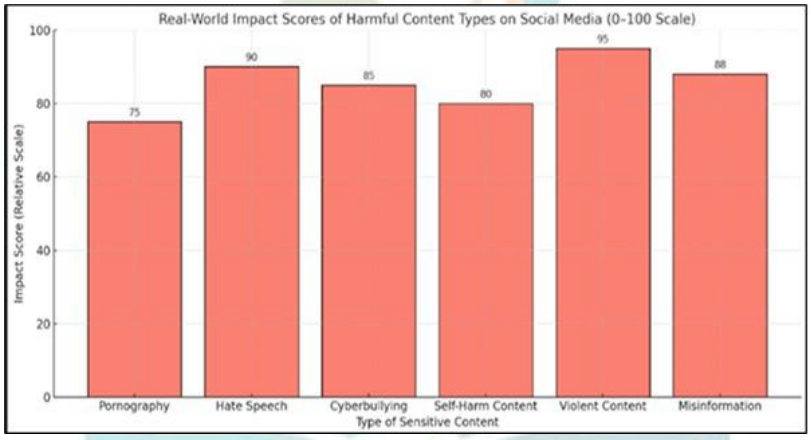

Figure 2: showing the graphics of the Types of sensitive content on social media and their impacts.

5. Challenges in AI moderation

5.1. Context Understanding Issues: AI struggles with sarcasm, hidden meanings, and multilingual content.
5.2. Adversarial Attacks: Manipulated text and images bypass AI filters.
5.3. Bias in AI Models: AI systems exhibit racial, gender, and cultural biases.
5.4. Scalability Issues: Real-time moderation for billions of users is computationally expensive.
5.5. Privacy Concerns: AI moderation can violate user privacy by analyzing personal messages.
AI Content Moderation: Loopholes, Solutions & Unsolved Challenges as Table 1.

TABLE I: TABLE FOR SHOWING LOOPHOLES, ISSUES, EXISTING SOLUTIONS AND THEIR UNSOLVED CHALLENGES

Loophole	Issue	Existing Solutions	Unsolved Problems
Context Understanding	AI struggles to understand sarcasm, hidden meanings, and local slang.	NLP models like BERT and T5 have improved contextual understanding.	Still difficult to moderate content in multi-lingual and low-resource languages.

Adversarial Attacks	Users bypass AI using misspellings, encoded text, or image manipulation.	Adversarial training and development of robust models are ongoing.	Deepfake text and images can still easily mislead AI systems.
Bias in AI Moderation	Racial, gender, and cultural biases in AI lead to unequal censorship.	Bias reduction and fairness-aware training techniques are being applied.	Completely eliminating bias from AI is still not possible.
Scalability Issues	Real-time moderation on large platforms is expensive and challenging.	Federated learning and edge computing solutions have been proposed.	Content moderation on decentralized platforms remains a significant challenge.
Privacy Concerns	AI often needs to analyse private messages, raising user privacy issues.	Techniques like homomorphic encryption and differential privacy are in use.	Effectively training AI without accessing personal user data is still not possible.
Lack of Explainability	Users don't know why their content is removed—AI decisions lack clarity.	Research on Explainable AI (XAI) and visualization frameworks is ongoing.	Achieving both high accuracy and explainability simultaneously remains an open issue.

6. Impact on victims and society

Unmoderated harmful content can lead to depression, anxiety, or suicide. A WHO 2023 report states that suicide is the 4th leading cause of death among 15–29-year-olds globally. Fake news during COVID-19 led to vaccine hesitancy and public unrest.

7. Existing Solutions and Their Gaps

- Facebook moderates 90% of hate speech before user reports as described in figure 3.
- YouTube auto-flags 70% of violative content
- Twitter fails during political events due to content volume

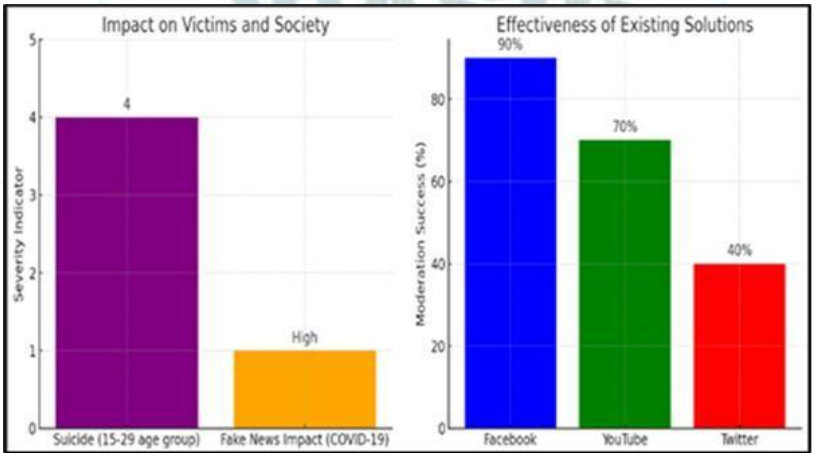

Figure 3: figure for showing the graphical images of Impacts and Effectiveness of Existing Solutions

7.1. Multi-Lingual & Low-Resource Content Moderation

Proposed Solution
Develop few-shot learning and transformer-based cross-lingual models that can perform accurate content moderation even with minimal training data in low-resource languages.
7.2. Adversarial Robustness Against Deepfakes & Manipulated Content
Proposed Solution:
Utilize self-supervised learning and contrastive learning techniques to help AI differentiate between real and fake content effectively.
7.3. Bias-Free AI Models for Fair Moderation
Proposed Solution:
Include diverse demographic groups in AI training datasets and implement fairness-aware adversarial training to reduce bias in moderation decisions.
7.4. Decentralized Content Moderation for Blockchain-Based Social Media
Proposed Solution:
Apply Federated Learning and Zero-Knowledge Proofs to ensure data privacy while enabling effective moderation on decentralized platforms.
7.5. Explainability in AI Moderation Decisions
Proposed Solution:
Implement Explainable AI (XAI) techniques that provide users with clear reasons behind the restriction or removal of their content.

8. Future Scope and Benefits
8.1. Enhanced Accuracy
Future models will be more context-aware and capable of understanding sarcasm, regional languages, and nuanced content with higher precision.
8.2. Real-time Moderation at Scale
AI systems will evolve to handle moderation across billions of posts and multimedia content instantly without human bottlenecks.
8.3. Cross-Platform Protection
Unified AI models could offer consistent moderation across multiple platforms like Facebook, Instagram, YouTube, and TikTok, providing safer ecosystems.
8.4. Protection of Mental Health
By reducing exposure to harmful content, AI moderation can help improve users' mental well-being and lower incidents of online harassment.
8.5. Customizable User Controls
Users might be free to set their own content preferences using AI, creating more personalized and safe content.
8.6. Support for Law
Enforcement Moderation systems can assist in flagging and reporting illegal content like child exploitation or terrorism, helping authorities act swift.

9. Case Studies
A. Self-Harm or Suicide-Related Content
Case Study: Molly Russell's Suicide and Instagram's Policy Changes
In 2017, 14-year-old Molly Russell from the UK died by suicide after engaging with graphic self-harm content on Instagram. Her family discovered distressing material about depression and suicide on her account posthumously. This incident prompted widespread criticism of Instagram's content policies. In response, Instagram announced in February 2019 that it would ban all graphic self-harm images, acknowledging that the platform was not where it needed to be regarding self-harm and suicide content. The platform committed to removing such content and making non-graphic self-harm content less discoverable.
B. Bulli Bai App Incident (2022)
In January 2022, an app named 'Bulli Bai' was discovered on the GitHub platform, displaying doctored images of Muslim women, including journalists and activists, for a mock online auction without their consent. This incident led to national outrage, legal actions against the app's creators, and intensified discussions about the safety and dignity of women in digital spaces.
C. Ranveer Allahbadia's Controversial Remarks (2025)
In February 2025, Indian YouTuber Ranveer Allahbadia faced significant backlash after making an obscene remark

during a YouTube comedy show. The incident sparked debates about content regulation on digital platforms, freedom of speech, and the responsibilities of content creators. Legal actions were initiated, and the case underscored the evolving landscape of digital content governance in India.

D. Manav Singh Suicide Case – Gurugram (2020)

In May 2020, 17-year-old Manav Singh from Gurugram died by suicide after being accused of sexual harassment by a classmate on Instagram. The allegations spread rapidly on social media, leading to online bullying and emotional trauma. Investigations later revealed inconsistencies in the accusations, sparking debates on digital vigilantism, mental health, and the need for responsible online behaviour.

E. Pollachi Sexual Assault Case – Tamil Nadu (2019)

In 2019, a group of men in Pollachi, Tamil Nadu, used social media platforms like Facebook to befriend young women (some minors), lured them into private locations, sexually assaulted them, and blackmailed them using recorded videos. The case involved over 200 victims and revealed how predators used digital tools for coordinated sexual exploitation.

F. Delhi Teen Murdered Over Instagram Abuse – Delhi (2021)

In December 2021, a 17-year-old boy was murdered in Delhi's Uttam Nagar area after he allegedly posted abusive comments on Instagram targeting a local gangster. The altercation escalated from online threats to real-world violence, ending in the fatal stabbing of the boy. This case shows how digital disputes can result in tragic offline consequences as described in figure 4.

G. Child Groomed via social media and Forced Marriage – West Bengal (2023)

In 2023, a 16-year-old girl from West Bengal was contacted by a man on Facebook who posed as a romantic partner. He later abducted and forced her into marriage. Investigations revealed that social media platforms are increasingly being used for child trafficking and exploitation under the guise of romantic relationships as described in figure 4.

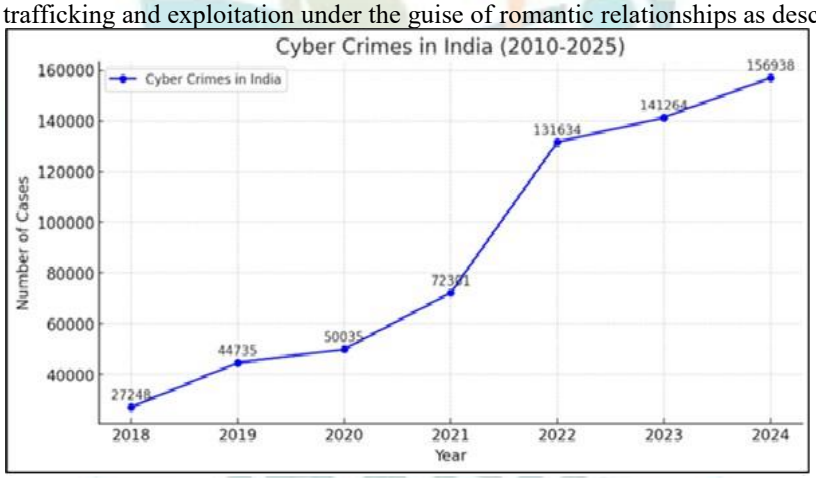

Figure 4: figure for showing crime rates in India year between 2010-2025

10. Conclusion

This paper has presented a comprehensive overview of AI-based content moderation techniques on social media platforms, addressing the increasing threats of hate speech, cyberbullying, misinformation, and explicit content. By leveraging machine learning, natural language processing (NLP), and computer vision models, AI provides powerful tools to identify and filter harmful content in real time. While current systems have shown promise, there are still challenges related to false positives, bias in algorithms, and adapting to evolving languages and behaviours. A multi-layered moderation approach combining AI with human oversight is essential to ensure accuracy and fairness. Ultimately, AI-based moderation systems represent a crucial step forward in creating a safer and more respectful digital environment for all users.

References

[1]. T. Gillespie, "Content moderation, AI, and the question of scale," ResearchGate, 2020. [Online]. Available: https://www.researchgate.net/publication/343798653_Content_moderation_AI_and_the_question_of_scale

[2]. Feerst, "The Use of AI in Online Content Moderation," American Enterprise Institute, Sep. 2022. [Online]. Available: https://platforms.aei.org/wp-content/uploads/2022/09/The-Use-of-AI-in-Online-Content-Moderation.pdf

[3]. M. L. Khan and M. Mudassar, "Detection and moderation of detrimental content on social media using artificial intelligence: A systematic

review," Social Network Analysis and Mining, vol. 12, no. 1, pp. 1–15, 2022.
[4]. Chekkee, "The Role of AI in Improving Content Moderation in Social Media," 2023. [Online]. Available: https://chekkee.com/the-role-of-ai-in-improving-content-moderation-in-social-media/
[5]. V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, and C. Tan, "Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation," arXiv preprint arXiv:2204.11788, 2022. [Online]. Available: https://arxiv.org/abs/2204.11788
[6]. X. Wang, S. Koneru, P. N. Venkit, B. Frischmann, and S. Rajtmajer, "The Unappreciated Role of Intent in Algorithmic Moderation of Social Media Content," arXiv preprint arXiv:2405.11030, 2024. [Online]. Available: https://arxiv.org/abs/2405.11030
[7]. H. Axelsen, J. R. Jensen, S. Axelsen, V. Licht, and O. Ross, "Can AI Moderate Online Communities," arXiv preprint arXiv:2306.05122, 2023. [Online]. Available: https://arxiv.org/abs/2306.05122
[8]. Oversight Board, "Content Moderation in a New Era for AI and Automation," 2023. [Online]. Available: https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation/ For case studies//
[9]. Self-Harm or Suicide-Related Content – Molly Russell Case and Instagram Response
[10]. BBC, "Instagram bans graphic self-harm images after Molly Russell's death," BBC News, Feb. 7, 2019. [Online]. Available: https://www.bbc.com/news/technology-47159093
[11]. Cyberbullying and Harassment – Indigenous Affairs Campaigns in Australia
[12]. C. Heenan, "Uluru youth co-chairs warn against disinformation ahead of election," The Australian, Mar. 2025. [Online]. Available: https://www.theaustralian.com.au
[13]. "Bulli Bai case," Wikipedia, Jan. 2022- https://en.wikipedia.org/wiki/Bulli_Bai_case
[14]. "Outrage over Indian YouTuber Ranveer Allahbadia raises social media regulation concerns," Associated Press, Mar. 2025- 'India's Got Latent': Ranveer Allahbadia controversy raises social media regulation concerns | AP News
[15]. "Manav Singh", Wikipedia, May 2020. - https://en.wikipedia.org/wiki/Manav_Singh