Journal of Science and Technology
ISSN: 2456-5660 Volume 10, Issue 02 (Feb -2025)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2025.v10.i02.pp39- 50*

Empowering Safe Online Spaces: AI in Gender Violence Detection and Prevention

Rahul Manche [1], FNU Samaah [2], Tejaswini Bollikonda [3] and Praveen Kumar Myakala [4, *]

[1] Independent Researcher, Hyderabad, India
[2] Independent Researcher, Chicago, Illinois, USA
[3] Saint Louis University, St. Louis, MO 63103
[4] Independent Researcher, Dallas, Texas, USA

[*] Corresponding author: Praveen Kumar Myakala, Email: Praveen.K.Myakala@gmail.com

To Cite this Article
Rahul Manche , FNU Samaah , Tejaswini Bollikonda and Praveen Kumar Myakala [*]**"Empowering Safe Online Spaces: AI in Gender Violence Detection and Prevention"***Journal of Science and Technology, Vol. 10, Issue 02-Feb 2025, pp39-50*

Article Info
Received: 24-11-2024 Revised: 02-02-2025 Accepted: 10-02-2025 Published:20 -02-2025

ABSTRACT

Gender-based online violence (GBOV) is a pervasive issue that disproportionately impacts women and marginalized genders, leading to psychological distress, economic consequences, and restricted participation in digital spaces. This article examines how Artificial Intelligence (AI) offers innovative solutions to detect and mitigate GBOV through tools such as sentiment analysis, hate speech detection, image recognition, and behavioral analysis. AI-powered interventions have significantly enhanced the ability to identify harmful content, automate moderation, and empower victims by providing real-time safeguards. Despite these advancements, challenges such as algorithmic bias, privacy concerns, and the evolving tactics of online abuse remain critical obstacles. This study highlights the importance of developing ethical AI systems, fostering multi-stakeholder collaboration, and implementing robust regulatory frameworks to create safer, more equitable online environments for all.

KEYWORDS: *Gender-based online violence (GBOV), Artificial intelligence (AI), Hate speech detection, Content moderation, Algorithmic bias*

I. INTRODUCTION

Gender-based online violence (GBOV) refers to a spectrum of abusive behaviors in digital spaces, including harassment, doxing, cyberstalking, and the non-consensual sharing of intimate content. These acts disproportionately target women and marginalized genders, infringing on their ability to engage freely and safely in online platforms. The impact of GBOV extends beyond the digital realm, resulting in profound psychological distress, erosion of social connections, and significant economic losses. Victims frequently resort to self-censorship or withdrawal from online spaces to avoid further abuse, curtailing their access to professional opportunities, social networks, and vital information [7, 9]. In severe cases, online abuse has led to reputational damage, employment termination, and physical violence [4].

Traditional approaches to addressing GBOV have proven insufficient in managing the scale and complexity of this issue. Content moderation by human reviewers, while valuable, is hindered by overwhelming volumes of harmful material, leading to inconsistent enforcement and delayed interventions [11]. Furthermore, the emotional toll of constantly moderating toxic content contributes to burnout and

turnover among moderators, compromising the effectiveness of these efforts. Automated systems, such as keyword-based filters, offer faster detection but often fail to recognize nuanced or context-specific abuse, such as coded language or newly emerging forms of harassment [13]. These limitations underscore the need for innovative, scalable solutions capable of adapting to the dynamic nature of online abuse.

Artificial intelligence (AI) offers a transformative approach to tackling GBOV. By leveraging advancements in Natural Language Processing (NLP), image recognition, and behavioral analysis, AI systems can process vast datasets, identify harmful patterns, and intervene in real time. Unlike traditional methods, AI enables scalable and consistent detection of abuse while addressing the limitations of manual and keyword-based moderation. For example, platforms such as Twitter and Facebook have successfully employed AI-driven models to identify and mitigate hate speech and harassment, significantly reducing the visibility of abusive content [7, 10].

II. AI-Powered Tools for Detecting and Preventing GBOV

Advancements in Artificial Intelligence (AI) have enabled the development of sophisticated tools to combat gender-based online violence (GBOV). These tools leverage natural language processing (NLP), image recognition, behavioral analysis, and real-time moderation techniques to detect and mitigate abusive behavior on digital platforms effectively. This section explores these tools, highlighting their applications, benefits, and real-world examples.

Natural Language Processing (NLP)

Natural Language Processing (NLP) plays a pivotal role in detecting hate speech, abusive language, and threats in text-based communication. Sentiment analysis, keyword detection, and contextual understanding allow NLP systems to identify harmful content, including subtle and coded language, that traditional filters often miss [13, 15]. For example, transformers like BERT and GPT excel in analyzing linguistic context, enabling the detection of harassment across multiple languages and dialects.

Platforms such as Twitter use NLP to identify and flag tweets containing hate speech or harmful content, significantly reducing exposure to users. Flagged content is either removed automatically or escalated to human moderators for further review, ensuring a balance between automation and accuracy [12]. This capability not only accelerates the moderation process but also enhances the precision of harmful content detection.

Key Benefit: NLP enables more accurate identification of hate speech, even when disguised through subtle or coded language, ensuring harmful content is addressed promptly.

Conclusion: By enabling early detection of abusive text, NLP enhances the scalability and efficiency of content moderation on platforms with vast user bases.

Image Recognition

AI-powered image recognition systems are crucial for detecting harmful visual content, including explicit imagery shared without consent or symbols associated with hate groups. These systems utilize deep learning models, such as convolutional neural networks (CNNs), to analyze images and videos with high accuracy. Hashing technologies, like those implemented by StopNCII.org, create unique digital fingerprints of flagged content, preventing its redistribution across platforms [5].

For instance, StopNCII.org empowers victims by allowing them to submit intimate images for hashing, ensuring these images cannot be re-uploaded on participating platforms. Similarly, Instagram employs image recognition to automatically remove explicit content violating its community guidelines.

Key Benefit: Image recognition offers a robust layer of protection against the non-consensual sharing of intimate images, helping victims regain control over their digital presence.

Conclusion: This technology provides critical support in combating visual abuse, reducing the spread of harmful content while safeguarding user privacy and dignity.

Behavioral Analysis

Behavioral analysis tools monitor user activity patterns to detect signs of harassment, such as repeated targeting of individuals or the coordination of abusive campaigns. These systems analyze metadata, posting frequency, and network interactions to identify suspicious behavior. For example, platforms use machine learning models to detect troll farms and orchestrated harassment campaigns that evade traditional moderation techniques [7].

Reddit employs behavioral analysis to detect and ban accounts involved in mass harassment or coordinated abuse, ensuring a safer environment for its users. This approach reduces the burden on victims to report abuse, enabling proactive interventions.

Key Benefit: Behavioral analysis prevents escalations by identifying and addressing harmful patterns before they cause significant harm.

Conclusion: By tracking user behavior, these tools enhance platform accountability and provide timely interventions against coordinated abuse.

Real-Time Content Moderation

AI-driven content moderation systems offer scalable solutions for handling user-generated content in real time. These systems rely on supervised and unsupervised learning models to filter harmful posts and comments effectively. Platforms such as Meta and YouTube have implemented AI-power moderation systems that identify and remove harmful content before it reaches a wide audience. For example, YouTube reports that over 94% of harmful videos flagged by AI systems are removed before users view them [7, 10].

Real-time moderation reduces the exposure of harmful content, alleviates the workload on human moderators, and improves the user experience by creating safer digital environments.

Key Benefit: Real-time moderation ensures swift action against harmful content, minimizing its impact on victims and online communities.

Conclusion: By automating routine moderation tasks, these systems enable human moderators to focus on complex cases requiring contextual understanding.

Proactive Safeguards

Proactive AI-powered safeguards are designed to prevent harm before it occurs. These tools include conversational AI chatbots, which monitor ongoing interactions, flag potentially abusive messages, and guide victims to support resources. For example, chatbots deployed on platforms like WhatsApp provide victims with immediate access to helplines, legal support, or psychological assistance.

Additionally, AI-powered customization tools, such as content filters and block lists, empower users to tailor their online experiences. These tools allow individuals to minimize exposure to harmful interactions by proactively managing their interactions and visibility.

Key Benefit: Proactive safeguards enable users to take control of their digital environments, reducing vulnerability to online abuse.

Conclusion: These tools provide an essential layer of prevention, empowering victims and enhancing the overall safety of digital platforms.

Multilingual and Context-Aware Capabilities

One of the most critical advancements in AI-powered tools is their ability to detect abusive behavior across languages and cultural contexts. Multilingual models trained on diverse datasets enable platforms to identify GBOV in underrepresented languages while adapting to cultural nuances [6, 12]. Facebook, for instance, uses AI to moderate content in over 50 languages, ensuring that non-English-speaking users receive the same level of protection as English speakers.

By reducing disparities in content moderation, multilingual tools address regional and cultural gaps, ensuring equitable protection for all users.

Key Benefit: Multilingual and context-aware tools bridge the gap in content moderation across diverse linguistic and cultural contexts, ensuring global inclusivity.

Conclusion: These capabilities enhance the effectiveness of AI tools in addressing GBOV worldwide, particularly in regions with limited prior access to moderation technologies.

AI-powered tools provide transformative capabilities to detect and prevent GBOV. From NLP and image recognition to real-time moderation and multilingual support, these technologies offer scalable and proactive solutions to combat online abuse. However, the effective implementation of these tools depends on continued innovation, ethical deployment, and collaboration among stakeholders. The next section delves into case studies that highlight the real-world impact of these technologies.

III. Case Studies and Success Stories

The implementation of AI in combating gender-based online violence (GBOV) has yielded promising results across multiple platforms and initiatives. This section highlights key examples that demonstrate the versatility and impact of AI-powered tools in reducing online harassment and fostering safer digital environments.

Meta's AI Moderation System

Meta (formerly Facebook) has employed advanced AI algorithms to detect and remove abusive content at scale. These algorithms utilize Natural Language Processing (NLP) and image recognition tools to identify hate speech, cyberstalking, and harmful imagery across various languages and cultural contexts. According to Meta's Transparency Report, over 95% of flagged harmful content is proactively detected before user reports, significantly reducing victims' exposure to abuse [10].

In 2022 alone, Meta's AI-driven tools helped remove over 30 million posts related to hate speech from Facebook and Instagram. These systems operate in real time, providing immediate interventions while generating detailed reports to assist human moderators with complex cases [7]. This combination of automation and human oversight ensures greater scalability and consistency in content moderation.

Takeaway: Meta's AI tools exemplify how large platforms can leverage AI to proactively address GBOV, reducing harmful content visibility and enabling safer user experiences.

StopNCII.org: Preventing Non-Consensual Content Sharing

StopNCII.org, a UK-based initiative by the NGO Revenge Porn Helpline, employs AI-driven hashing technology to combat the non-consensual sharing of intimate images. The platform enables victims to submit images, which are converted into unique digital fingerprints (hashes) without storing the original files. Social media platforms like Facebook and Instagram use these hashes to identify and prevent the re-upload of flagged content, giving victims control over their privacy [5].

This system has successfully prevented the redistribution of thousands of sensitive images, showcasing the potential of AI in empowering individuals against GBOV while preserving privacy and ensuring accountability.

Takeaway: By leveraging hashing technology, StopNCII.org provides a powerful, privacy-preserving solution to protect victims of image-based abuse.

Figure 1 Flowchart illustrates the hashing process in image-based abuse prevention. It shows the step-by-step workflow, starting from victim image upload to content blocking, with key processes such as hashing, secure storage, content matching, and automatic prevention.

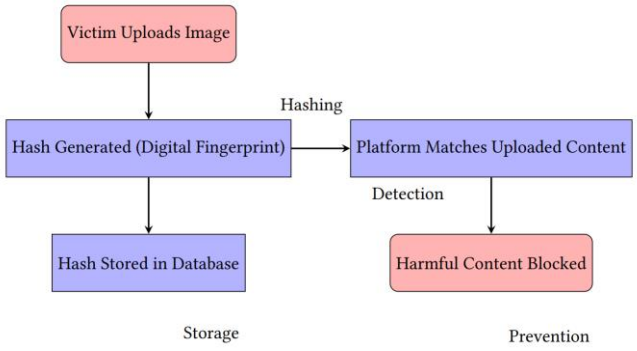

Figure 1: Flowchart showing the hashing process and its role in image-based abuse prevention.

Tackling Hate Speech on Twitter

Twitter has deployed AI models to identify and mitigate hate speech and harassment. These models utilize sentiment analysis and contextual understanding to detect harmful tweets, including coded language and regional variations [12]. In recent updates, Twitter reported that its AI systems proactively detect 65% of abusive content, significantly reducing reliance on user reports.

Moreover, collaborations with NGOs have enabled Twitter to refine its systems to address region-specific challenges, such as hate speech in underrepresented languages. These efforts underscore the importance of culturally adaptive AI tools in global content moderation.

Takeaway: Twitter's AI-based hate speech detection demonstrates the value of proactive moderation and the role of collaboration in improving platform accountability.

Proactive Intervention: Harm Prevention with AI

One of the most impactful uses of AI in combating GBOV is its ability to intervene in escalating situations before harm occurs. For example, AI-powered chatbots deployed on platforms like WhatsApp provide real-time support to victims of abuse. These chatbots can recognize signs of distress in conversations and offer resources such as helpline numbers, legal advice, or psychological counseling.

Additionally, predictive algorithms are being developed to monitor patterns of escalating online behavior, such as sustained targeting or coordinated harassment. Platforms like Reddit have piloted tools that flag accounts exhibiting harmful trends, enabling moderators to take preventive action before abuse escalates.

Takeaway: Proactive intervention tools showcase AI's ability to prevent harm by offering real-time support and addressing abusive behavior before it intensifies.

Localized Solutions by NGOs

Localized initiatives led by NGOs are essential in addressing region-specific challenges in GBOV. For instance, the SafeguardHER program in South Asia uses multilingual AI models to detect abusive language and imagery in regional dialects. These systems are integrated with support mechanisms, such as AI-powered chatbots, that provide victims with resources and guidance.

Similarly, the Middle East-based SMEX organization has developed tools to detect online harassment in Arabic, tackling the unique challenges posed by dialectal variations and context-specific abuse. These projects emphasize the need for culturally sensitive AI systems to ensure inclusivity and effectiveness.

Takeaway: Localized solutions highlight the importance of training AI models on diverse datasets and collaborating with community organizations to address GBOV effectively.

Impact on Specific Marginalized Groups

AI-powered interventions have proven particularly beneficial for marginalized groups, such as women of color, LGBTQ+ individuals, and people with disabilities. For instance, platforms that integrate AI-driven moderation tools have reduced the visibility of targeted harassment campaigns, which disproportionately affect these communities. Tools like image recognition have empowered victims of image-based abuse, while NLP models trained on inclusive datasets better detect slurs and discriminatory language [6].

Twitter has reported an increase in user satisfaction among LGBTQ+ users following the deployment of its hate speech detection tools. Similarly, NGO-led programs like SafeguardHER have provided critical support to women in rural areas, where access to digital safety resources is limited.

Takeaway: AI tools promote equity by reducing barriers to online participation for marginalized groups and addressing intersecting forms of discrimination.

These case studies illustrate the transformative potential of AI in combating GBOV across diverse platforms and regions. From largescale implementations by Meta and Twitter to localized efforts by NGOs like SMEX, AI-powered tools are creating safer and more inclusive digital environments. The examples of proactive intervention and targeted support for marginalized groups underscore the importance of inclusive and adaptive AI development. However, addressing challenges such as algorithmic bias, privacy concerns, and evolving abuse tactics remain crucial to ensuring the long-term effectiveness of these interventions.

IV. Challenges and Ethical Considerations

Despite the potential of Artificial Intelligence (AI) to mitigate gender based online violence (GBOV), its deployment poses significant challenges. Ethical considerations must be prioritized to ensure that AI systems are equitable, effective, and do not inadvertently exacerbate harm. This section explores key challenges, including algorithmic bias, privacy concerns, and the dynamic nature of online abuse.

Algorithmic Bias and Discrimination

AI systems are only as unbiased as the data on which they are trained. Many datasets used to train AI models reflect societal biases, including those related to gender, race, and language [6]. Consequently, AI tools may fail to detect GBOV directed at marginalized groups or may disproportionately flag content from underrepresented communities.

For example, research has shown that hate speech detection models often struggle with dialectal variations, such as African American Vernacular English (AAVE), leading to higher false-positive rates for non-abusive content [14]. These biases can undermine trust in AI systems and perpetuate existing inequalities.

Mitigating algorithmic bias requires the use of diverse and representative datasets, continual testing, and transparency in AI model development. Collaboration with communities affected by GBOV is also critical to ensure that AI tools address their unique needs and perspectives.

Privacy Concerns

AI systems often rely on extensive data collection to function effectively. For instance, behavioral analysis tools require access to user activity, while image recognition systems need datasets of flagged content for

training. Such practices raise concerns about data privacy and potential misuse, particularly when dealing with sensitive information like non-consensual images.

Ensuring privacy involves implementing robust data protection measures, such as encryption and anonymization, and adhering to legal frameworks like the General Data Protection Regulation (GDPR). Additionally, platforms should adopt a privacy-by-design approach, minimizing data collection to only what is necessary for GBOV prevention.

The Dynamic Nature of Online Abuse

Perpetrators of online abuse continually adapt their tactics to evade detection. For instance, the use of coded language, memes, and evolving platforms complicates the ability of AI systems to identify harmful behavior. Additionally, emerging technologies like deepfakes have introduced new forms of GBOV, such as manipulated videos used for blackmail or harassment.

Addressing these challenges requires AI models that are adaptive and capable of learning from new data in real time. Regular updates, alongside collaboration with experts in online abuse, can ensure that detection systems remain effective against evolving threats.

Balancing Free Speech and Safety

The use of AI for content moderation raises concerns about over censorship and its impact on free speech. Automated systems may inadvertently flag legitimate content, such as political commentary or satire, as harmful. This is particularly problematic in regions where online spaces are crucial for free expression and activism.

Striking a balance between safety and free speech involves refining AI algorithms to better understand context and incorporating human oversight into moderation processes. Transparent appeals mechanisms can also help users challenge wrongful content removal, fostering greater accountability.

Ethical AI Development and Governance

The ethical deployment of AI requires clear guidelines and governance structures. Governments, tech companies, and NGOs must collaborate to establish standards for transparency, accountability, and inclusiveness in AI development. Ethical AI frameworks, such as those proposed by UNESCO and the European Commission, emphasize principles like fairness, explainability, and human-centered design [17].

In addition, fostering public awareness about the capabilities and limitations of AI is essential to build trust and encourage responsible usage. By involving stakeholders at every stage of AI development, from design to deployment, the risks associated with these technologies can be mitigated.

The challenges and ethical considerations outlined above highlight the complexity of deploying AI to combat GBOV. While these technologies hold significant promise, addressing issues of bias, privacy, and evolving abuse tactics is essential to ensure their effectiveness and fairness. The next section examines policy frameworks and collaborative efforts necessary for fostering ethical and impactful AI deployment.

V. Policy Frameworks and Multi-Stakeholder Collaboration

The effective deployment of Artificial Intelligence (AI) to combat gender-based online violence (GBOV) requires comprehensive policy frameworks and coordinated efforts among governments, tech companies, NGOs, and civil society. This section outlines the role of policy in regulating AI, the importance of collaboration across stakeholders, and the need for public awareness and digital literacy initiatives.

Regulatory Frameworks for AI and GBOV

Governments play a critical role in establishing legal and ethical standards for AI development and deployment. Regulatory frameworks should focus on ensuring transparency, accountability, and fairness in AI systems. For example, the European Union's proposed AI Act introduces a risk-based classification system for AI applications, requiring rigorous oversight for systems that impact fundamental rights, including those used for content moderation [1], [18], [19].

National-level legislation, such as Australia's **eSafety** Act, has introduced mechanisms for holding platforms accountable for online harms, including GBOV. These policies mandate faster content removal, impose penalties for non-compliance, and encourage the adoption of AI for proactive moderation [3]. However, gaps in enforcement and jurisdictional differences remain challenges that require international cooperation.

Takeaway: Regulatory frameworks like the EU AI Act and eSafety Act demonstrate the importance of structured oversight and accountability in deploying AI for content moderation.

Table 1 shows a comparison of global AI regulations related to content moderation across different regions. It highlights key features such as transparency requirements, accountability mechanisms, and unique regional focuses, including privacy protections, algorithmic fairness, and proactive content moderation mandates.

Collaboration Between Stakeholders

No single entity can effectively address GBOV alone. Collaboration among tech companies, NGOs, and researchers is essential for developing inclusive and scalable solutions. Tech companies possess the resources to deploy AI systems, while NGOs and civil society organizations bring valuable insights into the needs and experiences of marginalized communities.

For instance, partnerships between platforms like Twitter and NGOs such as WITNESS have facilitated the development of AI tools that address region-specific GBOV challenges, such as identifying abusive content in underrepresented languages [8]. Similarly, industry coalitions like the Global Internet Forum to

Table 1: Global AI Regulations Related to Content Moderation

Region/Country	Regulation	Key Features
European Union	AI Act (Proposed)	• Risk-based classification of AI systems (e.g., "high-risk" for content moderation). • Transparency requirements for algorithmic decision-making. • Penalties for non-compliance: up to €30 million or 6% of global revenue.
United States	Algorithmic Accountability Act (Proposed)	• Requires impact assessments for bias, discrimination, and privacy risks. • Transparency mandates for automated decision-making processes. • Focus on safeguarding civil rights and free speech.
Australia	eSafety Act (2021)	• Platforms must remove harmful content within 24 hours of notification. • Encourages AI for real-time moderation. • Annual reports on moderation effectiveness required. • Penalties up to $555,000 AUD for non-compliance.
Canada	Digital Charter Implementation Act (Proposed)	• Algorithmic transparency and privacy protections. • User rights to contest automated decisions (e.g., content removal). • Focus on ethical AI for content curation and moderation.
China	Internet Information Service Algorithmic Recommendation Management Regulations (2022)	• Algorithms must align with "socialist core values." • Platforms must disclose recommendation logic. • Users can manage or disable algorithmic recommendations.
United Kingdom	Online Safety Bill (Proposed)	• Mandates proactive measures to prevent harmful content. • Encourages AI-driven content moderation with human oversight. • Transparency reports on content moderation required.

Counter Terrorism (GIFCT) demonstrate the potential for collaborative frameworks to tackle complex online harms.

Takeaway: Cross-sector collaboration is crucial for leveraging diverse expertise and ensuring AI tools are inclusive and culturally relevant.

Digital Literacy and Public Awareness Campaigns

Public awareness and education are critical components of a comprehensive strategy to combat GBOV. Digital literacy programs equip users with the knowledge to recognize and respond to online abuse, while also fostering responsible platform use. For example, UNESCO's Media and Information Literacy framework provides educational resources to empower individuals to navigate online spaces safely and effectively [16].

In addition, campaigns in the United States leverage multimedia outreach to raise awareness of GBOV and promote bystander intervention. Such initiatives complement AI tools by addressing the social and cultural dimensions of online abuse.

Takeaway: Digital literacy programs amplify the impact of AI tools by educating users and fostering a culture of online accountability.

Figure 2 illustrates the relative importance of various components within public awareness campaigns designed to combat GBOV. The largest segment, "Public Awareness Campaigns" at 36%, represents the overarching impact of these initiatives. This highlights the foundational role of raising awareness about GBOV as a precursor to other actions.

Incentivizing Platform Accountability

Policymakers must implement mechanisms that hold platforms accountable for their role in addressing GBOV. Transparency requirements, such as regular reporting on content moderation metrics, can incentivize companies to improve their AI systems. For instance, Meta's Transparency Report outlines the number of harmful posts removed, detection rates, and areas for improvement [10].

Additionally, tax incentives or funding programs can encourage smaller platforms and startups to invest in ethical AI development.

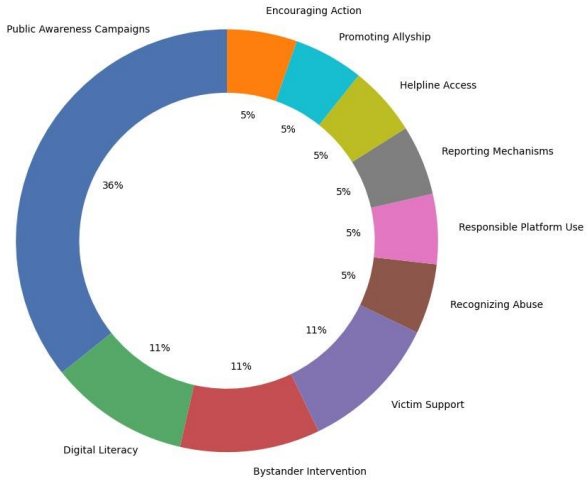

Figure 2: Role of Public Awareness Campaigns in Combating GBOV

Such measures ensure that even resource-constrained platforms contribute to creating safer online environments.

Takeaway: Incentivizing transparency and ethical AI development ensures that platforms of all sizes contribute to reducing GBOV.

Global Cooperation for Inclusive Solutions

GBOV is a global issue, requiring international cooperation to develop solutions that are culturally sensitive and contextually appropriate. Initiatives like the United Nations' Broadband Commission for Sustainable Development advocate universal standards in online safety, emphasizing the importance of gender equality in digital spaces [2].

Global partnerships must also address disparities in AI deployment. For example, developing regions often lack access to advanced moderation technologies, leaving communities vulnerable to GBOV. Collaborative funding and technology-sharing programs can bridge these gaps and ensure equitable access to protective measures.

Takeaway: Global cooperation ensures that AI tools for combating GBOV are accessible and effective across diverse cultural and economic contexts.

Policy frameworks and multi-stakeholder collaboration are essential to harnessing the full potential of AI in combating GBOV. By aligning technological innovation with ethical standards, fostering cross-sector partnerships, and promoting digital literacy, stakeholders can create safer and more inclusive online spaces. The next section will summarize key findings and emphasize the path forward for ethical and impactful AI deployment.

VI. Conclusion

Gender-based online violence (GBOV) is a pervasive and complex issue that undermines the safety, participation, and well-being of individuals in digital spaces, disproportionately affecting women and marginalized groups. This persistent challenge not only limits digital engagement but also leads to psychological distress, reputational damage, and economic harm. Traditional approaches, such as manual content moderation and user reporting, have proven insufficient to address the scale, complexity, and evolving nature of online abuse.

Artificial Intelligence (AI) has emerged as a transformative tool in combating GBOV, offering scalable and proactive solutions that surpass the limitations of conventional methods. Technologies like natural language processing (NLP), image recognition, behavioral analysis, and real-time content moderation enable platforms to detect and mitigate harmful behavior effectively. Case studies from major platforms, including Meta, Twitter, and YouTube, highlight the potential of AI-powered tools to reduce the visibility of harmful content and foster safer digital environments. Furthermore, localized initiatives like Chayn in South Asia and SMEX in the Middle East demonstrate the importance of culturally sensitive AI applications in addressing region-specific challenges.

However, the deployment of AI also presents significant challenges. Algorithmic bias risks excluding or harming marginalized groups, while privacy concerns regarding data collection and storage raise ethical questions. Additionally, the dynamic nature of online abuse, including emerging threats like deep-fake technology, necessitates continual adaptation and innovation. Balancing safety with free speech remains a critical concern, as over-censorship by automated systems can undermine the principles of free expression.

To address these challenges, a holistic and ethical approach is required. Regulatory frameworks, such as the EU AI Act and Australia's eSafety Act, provide essential guidelines to ensure transparency, accountability, and fairness in AI systems. Collaborative efforts among governments, tech companies, NGOs, and civil society are vital for developing inclusive and adaptable solutions. Public awareness

campaigns and digital literacy initiatives further empower users, equipping them to recognize and respond to GBOV while fostering a culture of accountability and respect in digital spaces.

Looking ahead, ongoing research and innovation are necessary to refine AI tools and address emerging forms of GBOV. Stakeholders must prioritize the development of adaptive and context-aware AI systems while ensuring inclusivity and fairness. Collaborative frameworks should aim to bridge disparities in access to technology, enabling equitable protection for vulnerable communities across the globe.

The potential of AI to combat GBOV is immense, offering opportunities to create safer and more inclusive online environments. Its success, however, depends on collective action, ethical implementation, and a shared commitment to continuous improvement. By uniting technological innovation with human-centered values, we can make meaningful progress in safeguarding digital spaces for all users.

Acknowledgments

The authors express their gratitude to the research community for their foundational work in gender bias detection, which served as an inspiration for this study. Special thanks are extended to organizations and initiatives actively combating gender-based online violence, whose insights and case studies have significantly enriched the understanding of this complex issue. Additionally, the authors acknowledge the contributions of interdisciplinary teams working at the intersection of technology, ethics, and social justice, whose efforts continue to shape equitable and inclusive AI systems.

References

[1] European Commission. 2021. Proposal for a Regulation Laying Down Harmonized Rules on Artificial Intelligence (AI Act). *https://eur-lex.europa.eu/legal-content/ EN/TXT/?uri=CELEX:52021PC0206*

[2] United Nations Broadband Commission. 2021. Towards a Resilient, Safe, and Inclusive Digital Future. *https://broadbandcommission.org*

[3] Australian eSafety Commissioner. 2021. Australia's eSafety Act: Enhancing Online Safety. *https://www.esafety.gov.au*

[4] Nicola Henry, Asher Flynn, and Anastasia Powell. 2020. Technology-facilitated domestic and sexual violence: A review. Violence against women 26, 15-16 (2020), 1828–1854. *https://doi.org/10.1177/1077801219875821*

[5] StopNCII.org Initiative. 2021. StopNCII.org: An AI-driven solution for combating image-based abuse. *https://stopncii.org*

[6] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35. *https://doi.org/10.1145/3457607*

[7] The Economist Newspaper. 2021. Measuring the prevalence of online violence against women. *https://onlineviolencewomen.eiu.com*

[8] WITNESS Organization. 2022. WITNESS: Addressing Online Harms through AI and Collaboration. *https://witness.org*

[9] World Health Organization. 2021. Violence Against Women Prevalence Estimates. *https://www.who.int/publications/i/item/violence-against-women*

[10] Meta AI Research. 2022. Our New AI System to Help Tackle Harmful Content. https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmfulcontent

[11] Sarah T. Roberts. 2019. Behind the Screen: Content Moderation in the Shadows of Social Media. Yale University Press. *https://doi.org/10.2307/j.ctvhrcz0v*

[12] Sayar Ghosh Roy, Ujwal Narayan, Tathagata Raha, Zubair Abid, and Vasudeva Varma. 2021. Leveraging multilingual transformers for hate speech detection. arXiv preprint arXiv:2101.03207 (2021). *https://doi.org/10.48550/arXiv.2101.03207*

[13] Raiswa Saha, Sakshi Ahlawat, Umair Akram, Uttara Jangbahadur, Amol S Dhaigude, Pooja Sharma, and Sarika Kumar. 2024. Online abuse: a systematic literature review and future research agenda. International Journal of Conflict Management (2024). *https://doi.org/10.1108/IJCMA-09-2023-0188*

[14] Maarten Sap et al. 2020. The Risk of Racial Bias in Hate Speech Detection. Proceedings of the ACL (2020),1668–1679. *https://doi.org/10.18653/v1/P19-1163*

[15] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. ACL Anthology (2017), 1–10. *https://doi.org/ 10.18653/v1/W17-1101*

[16] UNESCO. 2022. Media and Information Literacy: UNESCO's Framework for Digital Literacy. *https://www.unesco.org/en/media-information-literacy*

[17] UNESCO. 2022. UNESCO Guidelines for Ethical AI Development. *https://unesdoc.unesco.org/ai-guidelines*

[18] Myakala, P. K., Jonnalagadda, A. K., & Bura, C. (2025). The Human Factor in Explainable AI Frameworks for User Trust and Cognitive Alignment. *https://dx.doi.org/10.2139/ssrn.5103067*

[19] Myakala, P. K. (2024). Consciousness in Machines: A Critical Exploration. International Journal of Multidisciplinary Research and Analysis, 7(12), 10-47191. *https://doi.org/10.47191/ijmra/v7-i12-18*