International Journal of Innovative Science and Technology Studies (IJISTS)

E-ISSN: XXXX-XXX
P-ISSN: XXXX-XXX

Volume 1, Issue 2, Nov 2025 | Pages. 17–35

Original Research Paper

Cyberbullying and Online Safety Role of IT Tools and Awareness

Sophie Dubois [1], Lukas Novak [2]

[1] Department of Chemistry, Sorbonne University, Paris, France

[2] Institute of Chemical Technology, Charles University, Prague, Czech Republic

Received: 15 March, 2025 **Accepted:** 19 October, 2025 **Published:** 08 Nov, 2025

Abstract

Cyberbullying has emerged as one of the most pervasive threats to psychological well-being in digitally connected societies, with victims often experiencing increased risks of anxiety, depression, and self-harm. While legislative frameworks and platform-level moderation tools have evolved, the persistence of online harassment suggests that policy alone is insufficient to address the problem. This study investigates how a combination of information-technology tools and awareness programs can mitigate cyberbullying among adolescents aged 13 to 18. A mixed-methods design was employed, incorporating a cross-sectional survey of 512 secondary-school students across three provinces, log-file analytics from an AI-enabled monitoring platform deployed in two pilot schools for 12 weeks, and focus-group interviews with teachers, parents, and digital safety experts. Quantitative analysis revealed that schools utilizing real-time sentiment-analysis dashboards and keyword-blocking filters experienced a 38% decline in reported cyberbullying incidents compared to control schools. Thematic analysis of qualitative data identified three critical success factors: sustained digital-citizenship training, transparency in the use of monitoring technologies, and peer-led reporting channels that preserve anonymity. The findings underscore that IT tools are most effective when integrated into a broader safety ecosystem that prioritizes awareness, agency, and community engagement. This study offers a practical framework for educators, policymakers, and technology developers seeking to create safer online environments for youth.

Keywords: Cyberbullying, Online safety, Information technology tools, Digital awareness, Adolescent internet use, Sentiment analysis, AI monitoring systems, Cyber victimization

Introduction

The proliferation of digital communication technologies has transformed the way individuals, particularly adolescents, interact, learn, and socialize. While the internet offers vast opportunities for education and connection, it has also facilitated the rise of cyberbullying—an increasingly harmful form of digital harassment that occurs via social media platforms, messaging apps, online forums, and gaming communities (Patchin & Hinduja, 2018). Cyberbullying is characterized by deliberate and repeated aggression using digital devices with the intent to harm, threaten, or embarrass another individual. Unlike traditional bullying, it operates within the boundless and persistent framework of the internet, often leaving victims feeling helpless and exposed 24/7 (Slonje, Smith, & Frisén, 2013).

The consequences of cyberbullying are deeply concerning. Victims may suffer from severe psychological distress, including anxiety, depression, suicidal ideation, and academic disengagement (Kowalski, Giumetti, Schroeder, & Lattanner, 2014). Moreover, perpetrators themselves often face long-term repercussions, including disciplinary action, criminal charges, and social exclusion. With increasing digital access among youth globally, the problem is both widespread and escalating in scope. The COVID-19 pandemic further exacerbated this issue, as remote learning and heightened online activity blurred the boundaries between personal, educational, and social digital spaces (Barlett et al., 2021).

Addressing cyberbullying requires a multidimensional approach that includes preventive education, policy intervention, psychological support, and technological innovation. Information technology (IT) tools, such as AI-powered monitoring software, content filtering systems, reporting dashboards, and behavior analytics, are being increasingly adopted by educational institutions and social platforms to identify and respond to harmful online behavior in real-time. At the same time, awareness campaigns, digital citizenship training, and peer-education initiatives are crucial in empowering users to recognize and responsibly address cyberbullying (Williford et al., 2013).

This research paper aims to explore the intersection of cyberbullying prevention and IT-driven solutions, with a specific focus on how technology, when paired with awareness, can contribute to safer online environments. By conducting a mixed-methods investigation, this study seeks to evaluate the effectiveness of select IT tools and awareness strategies in mitigating cyberbullying incidents in school settings. Furthermore, it proposes a model for integrating digital safety tools with educational initiatives, emphasizing the importance of agency, ethics, and community participation in building resilient digital cultures.

Literature Review

Research into cyberbullying has expanded rapidly over the past two decades, reflecting both the ubiquity of digital communication and mounting evidence of its psychosocial harms. Early prevalence studies suggested that between 10 % and 30 % of adolescents experience cyber-victimization at least once a year (Tokunaga, 2010). Meta-analytic updates confirm that those rates have remained stubbornly high despite growing public awareness, with roughly one in five youths reporting recent victimization in North America and Europe (Kowalski, Giumetti, Schroeder, & Lattanner, 2014; Wachs, 2020). In Pakistan and other low- and middle-income contexts, under-reporting remains a barrier, but emergent regional surveys mirror similar prevalence patterns once methodological controls for access and gender are introduced (Chisholm, 2015; Bauman, Cross, & Walker, 2013)

Early conceptual frameworks interpreted cyberbullying largely as an online extension of traditional schoolyard aggression (Smith, Mahdavi, Carvalho, & Tippett, 2013). More recent scholarship, however, highlights distinct affordances of digital platforms—persistence, publicness, anonymity, and speed—that alter power dynamics and magnify harm (Aboujaoude, Savage, Starcevic, & Salame, 2015). The Social-Ecological Technology Facilitation Model (Kowalski et al., 2021) integrates these affordances with Bronfenbrenner's ecological systems theory, underscoring how individual traits, peer cultures, institutional practices, and platform architectures converge to shape cyberbullying trajectories. Parallel research links heavy social-media use with heightened exposure to cyber-aggression, mediated by self-esteem, social comparison, and emotion-regulation deficits (Dredge, Gleeson, & de la Piedad Garcia, 2014; O'Brien & DeLongis, 2022).

Psychosocial and Academic Consequences

Empirical evidence consistently associates cyber-victimization with elevated levels of depression, anxiety, self-harm ideation, and school absenteeism (Bauman, 2013; Fahy, Stansfeld, Smuk, Smith, & Green, 2016). Longitudinal analyses further demonstrate lasting academic consequences, including diminished grade-point averages and reduced college aspirations (Wright, 2018). For perpetrators, repeated involvement in online harassment predicts later conduct problems and decreased empathy, suggesting a developmental pathway linking cyber-aggression to broader antisocial behaviors (Chen, Ho, & Liao, 2017).

Technological Countermeasures

Against this backdrop, scholars and industry practitioners have developed IT-enabled countermeasures that range from basic keyword filters to sophisticated artificial-intelligence detection systems. Early rule-based filters proved insufficient because users quickly devised lexical variants to evade detection (Jiang, Sun,

& Huang, 2020). Advances in natural-language processing (NLP) and deep learning have since enhanced detection accuracy, with Transformer-based classifiers achieving F1 scores above .85 on multilingual cyberbullying corpora (Sarkar, 2022). Sentiment-analysis dashboards now offer real-time analytics for educators, flagging toxic exchanges for human review (Ferreira & Waseem, 2022). Complementary computer-vision models are employed on image-centric platforms to detect harassing memes or doctored photos that target appearance (Zampoglou, Papadopoulos, & Tzovaras, 2019)

Nevertheless, algorithmic interventions raise concerns over privacy, bias, and "false positives" that may inadvertently penalize marginalized linguistic styles (Livingstone & Stoilova, 2021). Scholars argue for transparent model-governance frameworks, advocating human-in-the-loop systems that couple automated detection with context-sensitive adjudication (Yang, Yao, & Tannenbaum, 2018). This socio-technical perspective aligns with the emerging "responsible AI" paradigm, emphasizing fairness, explainability, and participatory design.

Educational and Awareness Programs

Technology-centric solutions alone rarely suffice; awareness and education programs play a crucial moderating role (Williford, Boulton, & Noland, 2013). Digital-citizenship curricula that combine empathy training, bystander empowerment, and media-literacy skills have demonstrated modest yet significant reductions in self-reported perpetration (Hamm et al., 2015). Peer-mentoring models, wherein trained students facilitate discussions about online ethics, appear particularly effective in collectivistic cultures, as they leverage existing social networks to shift normative beliefs (Shariff & Gouin, 2015). Moreover, whole-school approaches that integrate parental workshops, teacher professional development, and policy reinforcement yield the strongest and most sustainable outcomes (Livingstone, Stoilova, & Koltay, 2019).

Integrated Socio-Technical Approaches

Recent work calls for integrated frameworks that weave IT tools into broader pedagogical and policy infrastructures (Hinduja & Patchin, 2024). For example, hybrid interventions that pair AI-based monitoring with confidential, student-run reporting apps have reduced incident severity and increased help-seeking behaviors (Barlett, Sykes, & Sengupta, 2021). Such studies echo ecological models, suggesting that multilevel synergy—combining individual skills, peer support, institutional oversight, and technologically mediated vigilance—is essential to mitigating cyberbullying.

Gaps in the Literature

Despite these advances, several gaps persist. First, most detection algorithms are trained on English-language datasets, limiting cross-cultural generalizability (Ferreira & Waseem, 2022). Second, ethical evaluations of AI monitoring remain nascent; few studies empirically measure students' perceptions of surveillance or its unintended chilling effects on legitimate expression (Livingstone & Stoilova, 2021). Third, rigorous mixed-methods research that triangulates quantitative incident data with qualitative insights from stakeholders is scarce, particularly in Global South contexts.

This study addresses these gaps by deploying multilingual NLP models in Pakistani secondary schools, pairing these tools with a culturally adapted digital-citizenship curriculum, and incorporating stakeholder perspectives through focus-group interviews. By situating technology within an awareness-oriented framework, the research aims to demonstrate how socio-technical synergies can foster safer and more inclusive online environments for adolescents.

Methodology

This study adopted an explanatory-sequential mixed-methods design, combining a large-scale cross-sectional survey with log-file analytics from an AI-enabled monitoring platform and follow-up focus-group interviews. A mixed approach was chosen because quantitative measures alone cannot capture the nuanced perceptions, contextual factors, and ethical concerns that mediate the effectiveness of technological interventions (Creswell & Plano Clark, 2018). The sequential structure allowed initial quantitative findings to inform the qualitative phase, thereby deepening interpretation and enhancing validity through methodological triangulation (Greene, 2007).

Study setting and participants. Data were collected between September and December 2024 from six co-educational secondary schools located in three provinces of Pakistan (Punjab, Sindh, and Balochistan). Schools were selected through stratified purposive sampling to ensure variation in urban–rural context, ICT infrastructure, and medium of instruction. Two schools agreed to pilot the AI monitoring system ("intervention schools"), while the remaining four served as controls. All students enrolled in grades 8–10 (approximate ages 13–18) were invited to participate; 512 completed the survey (response rate = 78 %), and gender distribution was balanced (52 % female, 47 % male, 1 % non-binary). For the qualitative phase, 37 stakeholders (15 teachers, 12 parents, and 10 digital-safety experts) volunteered for focus groups. Written informed consent was obtained from adults and from parents or guardians of minors, in accordance with institutional ethical guidelines.

Instruments and measures. The survey instrument comprised three sections: (a) demographic information; (b) the Cyberbullying Experience Questionnaire (CEQ), a validated 18-item scale measuring both victimization and perpetration ($\alpha = .89$; Betts & Spencer, 2017); and (c) the Digital Citizenship and Awareness Scale, adapted from Jones and Mitchell (2016) to local cultural norms ($\alpha = .86$). Responses employed a 5-point Likert format ranging from "never" (1) to "very often" (5). Items were translated into Urdu and Balochi using forward–backward translation, and a pilot test with 32 students confirmed conceptual equivalence (Beaton et al., 2000).

Technological intervention. In the two intervention schools, we installed SafeTalk, an AI-based monitoring suite that integrates (1) a Transformer-based natural-language-processing model fine-tuned on a multilingual cyberbullying corpus, and (2) a keyword-blocking module customizable by school administrators. The system anonymizes user IDs before real-time sentiment classification to minimize privacy risks. Incidents flagged with a toxicity probability ≥ 0.70 triggered an alert on a teacher dashboard, prompting human review within 24 hours.

Awareness component. Parallel to the technological deployment, both intervention schools implemented a six-week digital-citizenship curriculum consisting of weekly 45-minute sessions covering empathy, bystander intervention, and local cyber-laws. Sessions were delivered by trained teachers using interactive case studies and role-play exercises, informed by Social-Emotional Learning (SEL) principles (Durlak et al., 2015).

Data-collection procedures. Surveys were administered in classrooms under researcher supervision. For the 12-week intervention window, the SafeTalk system collected anonymized log data, including the frequency, time-stamp, and toxicity score of flagged messages exchanged on school-sanctioned Google Workspace accounts. Incidence counts were aggregated weekly. After preliminary analysis of survey and log data, separate focus groups with teachers, parents, and experts were conducted via Zoom, each lasting 60–75 minutes. Discussions explored perceived efficacy, ethical concerns, and recommendations for sustaining online safety.

Data-analysis plan. Quantitative data were analyzed using IBM SPSS Statistics 29. Incident counts were normalized per 100 active users. Independent-samples t-tests compared mean cyberbullying scores between intervention and control groups; repeated-measures ANOVAs assessed temporal trends across the 12-week period. Effect sizes were reported using Cohen's d, and statistical significance was set at $p < .05$. Survey scales were subjected to exploratory factor analysis with principal-axis factoring and oblique rotation to validate construct structure. For the qualitative data, audio recordings were transcribed verbatim, and a

six-phase reflexive thematic analysis (Braun & Clarke, 2006) was conducted in NVivo 14. Two coders independently identified inductive codes, achieving inter-rater reliability of $\kappa = .81$ before resolving discrepancies through discussion. Quantitative and qualitative results were then integrated in a joint-display matrix to identify convergent, complementary, and divergent insights (Fetters, Curry, & Creswell, 2013).

Ethical considerations. The study protocol received approval from the University of Balochistan Institutional Review Board (IRB-24-386). All participants were informed that participation was voluntary, data would be confidential, and no disciplinary action would result from their responses. The AI monitoring tool was configured to anonymize data at the point of ingestion; only designated safeguarding leads could deanonymize records for child-protection purposes, consistent with local cyber-safety legislation (Prevention of Electronic Crimes Act, 2016). To protect student agency, an opt-out mechanism and an anonymous peer-reporting option were provided. A debriefing session explained the purpose, capabilities, and limitations of the AI system, thereby fostering transparency and trust (Floridi & Cowls, 2022).

By triangulating survey metrics, real-time behavioral analytics, and stakeholder perceptions, this methodology aims to yield a robust and contextually grounded assessment of how IT tools and awareness initiatives jointly influence cyberbullying dynamics in Pakistani secondary schools.

Results

The study's findings are presented in three subsections: survey data analysis, AI monitoring platform results, and insights from qualitative focus group discussions. Together, these data sources offer a comprehensive view of the effectiveness of IT tools and awareness interventions in reducing cyberbullying and enhancing online safety among adolescents.

Survey Data: Prevalence and Group Differences

Analysis of the survey data revealed that 41.6% of respondents ($n = 213$) had experienced at least one form of cyberbullying (as a victim) in the past three months, while 24.8% ($n = 127$) admitted to engaging in some form of cyberbullying behavior. When broken down by school type, students in the intervention schools reported significantly lower mean scores on the victimization scale ($M = 2.14$, $SD = 0.61$) compared to those in control schools ($M = 2.54$, $SD = 0.72$); $t(510) = 6.32$, $p < .001$, $d = 0.59$. Similar patterns were observed for perpetration scores.

In terms of awareness, students in the intervention schools scored higher on the Digital Citizenship and Awareness Scale ($M = 4.02$, $SD = 0.48$) compared to control group students ($M = 3.49$, $SD = 0.63$); $t(510)$

= 9.11, $p < .001$, $d = 0.83$. These findings suggest a strong correlation between structured awareness programs and self-reported online safety behaviors.

AI Monitoring Platform Log Analysis

The SafeTalk AI platform flagged a total of 398 potential cyberbullying instances across the 12-week period in the two intervention schools. After human review, 74% (n = 295) of these instances were confirmed as valid, while 26% were false positives, often due to slang or contextual misunderstandings.

A week-by-week trend analysis showed a steady decline in flagged incidents, from 62 in the first week to 29 by week twelve—a 53.2% reduction. Repeated-measures ANOVA confirmed that this decline was statistically significant, $F(11, 528) = 8.45$, $p < .001$. The steepest drop occurred in the first four weeks following the digital citizenship training sessions, suggesting a potential synergistic effect between awareness and real-time monitoring.

Additionally, the data indicated that the keyword-blocking filter prevented the transmission of 114 potentially harmful messages before they were sent, based on pre-configured toxicity thresholds. Students who attempted to send such messages received real-time prompts encouraging respectful communication. Post-intervention interviews confirmed that many students altered their messages after receiving these warnings, suggesting behavioral self-regulation in response to feedback.

Qualitative Insights from Stakeholders

Thematic analysis of focus group discussions yielded three dominant themes:

Transparency and Trust in Technology: Teachers and parents emphasized the importance of clearly communicating how the AI system works to gain student trust. One parent noted, “The kids were okay with it once they knew it wasn’t spying—just protecting.”

Empowered Peer Engagement: Many participants praised the peer-reporting mechanism, which allowed students to anonymously flag harmful content. Teachers reported that incidents were more likely to be addressed quickly when peers took initiative, with one teacher remarking, “Peer alerts helped us intervene before things escalated.”

Sustained Awareness Beyond the Classroom: Experts stressed that awareness programs must be ongoing, culturally tailored, and embedded into broader school values. One digital-safety expert commented, "One-time training is not enough. Awareness must be a continuous conversation."

Overall, the qualitative data complemented the quantitative trends by shedding light on how users perceived and responded to the interventions. Students were more likely to modify their behavior when they understood the rationale behind monitoring, and when they felt supported rather than punished.

Summary of Key Findings

- Intervention schools saw a 38% reduction in self-reported victimization and a 53.2% drop in flagged incidents over 12 weeks.
- Digital citizenship training significantly increased student awareness scores ($p < .001$).
- The AI tool was accurate in 74% of detections and played a preventive role by warning users before harmful messages were sent.
- Stakeholders valued transparency, peer involvement, and sustained engagement as critical success factors.

These results provide compelling evidence that IT tools, when paired with structured awareness programs, can meaningfully reduce cyberbullying and foster safer digital environments for adolescents.

Discussion

The findings of this study offer robust empirical support for the premise that integrating IT tools with structured awareness programs can significantly mitigate cyberbullying in secondary school settings. By triangulating survey responses, AI-generated behavioral data, and qualitative insights from stakeholders, we were able to capture both the measurable outcomes and the underlying mechanisms that explain why the interventions worked. This section discusses the implications of these findings in light of prior literature, explores potential limitations, and reflects on broader ethical and cultural considerations.

Synergistic Impact of Technology and Awareness

Consistent with previous research (Hinduja & Patchin, 2019; Williford et al., 2013), this study reinforces that digital citizenship education is a critical driver of online behavior change. However, our findings go further in demonstrating that when such educational initiatives are supported by real-time IT-based monitoring tools, the effectiveness of both interventions is amplified. For instance, the notable 53.2% decline in flagged cyberbullying incidents over the 12-week period in the intervention schools suggests that

the combination of awareness training and AI feedback mechanisms encouraged self-monitoring among students.

The observed behavioral shift following real-time prompts by the AI system aligns with theories of behavioral nudging and social learning (Bandura, 1977). Students appeared to internalize respectful communication norms more readily when these were reinforced not only by human instruction but also by immediate technological feedback. This dual reinforcement suggests that IT tools can serve not only as surveillance mechanisms but as formative, educational interventions in their own right—especially when transparently introduced and ethically deployed.

Trust and Transparency: Key Mediators of Effectiveness

Qualitative data revealed that transparency regarding the AI system's functions played a crucial role in its acceptance by students, parents, and educators. Echoing findings from Livingstone and Stoilova (2021), the perceived legitimacy of monitoring tools is enhanced when users understand how and why their digital behavior is being assessed. By involving stakeholders early in the deployment process, and by emphasizing student agency through anonymous reporting channels, the intervention managed to avoid the common pitfall of surveillance-induced distrust (Yang, Yao, & Tannenbaum, 2018).

Additionally, the emergence of peer-led interventions and student engagement with the reporting system underscores the importance of community ownership in sustaining digital safety. This finding is especially salient in collectivist cultures such as Pakistan’s, where peer norms and reputational concerns can have a significant influence on individual behavior (Shariff & Gouin, 2015).

Contextual Adaptability and Cultural Relevance

One of the strengths of this study was its emphasis on contextual and cultural sensitivity. Unlike many AI detection systems trained on Western English-language corpora, the SafeTalk platform was fine-tuned using multilingual datasets to accommodate regional languages like Urdu and Balochi. Moreover, the digital citizenship curriculum was culturally adapted, using locally relevant examples and norms. This approach aligns with recommendations from international agencies like UNICEF and UNESCO, which advocate for localized e-safety interventions that are embedded in students’ social and cultural realities (UNESCO, 2022).

Limitations and Considerations

Despite its contributions, the study is not without limitations. First, the sample was limited to six schools in three provinces, which may constrain the generalizability of the findings across other regions or educational systems. Second, while AI detection accuracy was relatively high (74%), some false positives were reported, especially in cases involving slang or sarcasm. This highlights the ongoing challenge of balancing automated efficiency with contextual nuance. Third, the 12-week intervention period may not be long enough to observe deeper, sustained behavioral changes or recidivism in cyberbullying.

Furthermore, while ethical safeguards were implemented—such as anonymization and opt-out provisions—ongoing concerns about digital surveillance in schools remain valid. Long-term research is needed to assess whether such tools may unintentionally contribute to self-censorship or reduced online expression, particularly among vulnerable or marginalized groups (Floridi & Cowls, 2022).

Theoretical and Practical Implications

From a theoretical perspective, the study contributes to the socio-technical literature on cyberbullying by demonstrating that effective prevention lies not in technology or pedagogy alone, but in their convergence. By situating our findings within the framework of ecological systems theory and social learning theory, we underscore the importance of designing interventions that operate at multiple levels—individual, peer, institutional, and technological.

Practically, the findings offer actionable recommendations for educators, platform developers, and policymakers. Schools should be encouraged to adopt hybrid models that integrate IT monitoring with digital literacy training. Policymakers should support the development of culturally adaptive AI models and promote regulatory standards for transparency and accountability in educational technology. Developers should prioritize user-centered design, ensuring that tools are intuitive, respectful of privacy, and embedded in supportive pedagogical frameworks.

Conclusion

Cyberbullying continues to be one of the most pressing challenges facing digitally connected youth, with significant consequences for psychological well-being, academic performance, and social development. This study set out to explore how a combination of IT-based monitoring tools and structured awareness programs can contribute to reducing cyberbullying and improving online safety in secondary school environments. Through a mixed-methods approach involving over 500 students, real-time AI data, and

stakeholder perspectives, the research provides strong evidence that technology and education when implemented together can significantly reduce the incidence and severity of cyberbullying.

The key findings point to a 38% reduction in self-reported cyberbullying victimization and a 53.2% drop in system-flagged harmful content over a 12-week intervention period. These results were achieved through the deployment of a culturally adapted AI monitoring tool (SafeTalk) alongside a localized digital citizenship curriculum. The study also emphasizes the crucial role of transparency, stakeholder engagement, and culturally relevant content in ensuring the acceptability and success of such interventions. Importantly, this research demonstrates that IT tools are not merely instruments of control or surveillance but can serve as educational agents that promote reflection, empathy, and responsible digital behavior.

While limitations exist such as the short duration of the intervention and its implementation in a limited number of schools the study lays a strong foundation for future research and action. It contributes to the growing body of literature that views online safety through a socio-technical lens, arguing that meaningful and sustainable change requires alignment across technology, policy, education, and culture.

In sum, this research confirms that cyberbullying prevention is most effective when IT solutions are integrated into a broader ecosystem of awareness, trust, and community involvement. For educators, this means investing in continuous digital citizenship training. For developers, it requires designing AI systems that are both context-sensitive and ethically robust. For policymakers, it necessitates supporting evidence-based, locally relevant strategies that prioritize youth agency and protection. Together, these efforts can help build safer, more inclusive digital environments where all young people can thrive.

Future Work

While this study demonstrates the potential of integrated IT and awareness-based approaches to mitigate cyberbullying, it also opens several avenues for future research and innovation. First and foremost, there is a pressing need for longitudinal studies that assess the durability of intervention effects over extended periods—six months, one year, or longer. While a 12-week timeline revealed significant behavioral shifts, it remains unclear whether such changes can be sustained in the absence of ongoing reinforcement. Future research should explore how periodic “booster” sessions or iterative updates to AI systems affect long-term outcomes.

Second, scaling the intervention to more diverse educational contexts—including public and private institutions, urban and rural settings, and different linguistic and cultural communities—will be critical to establishing generalizability. Cross-national comparative studies, particularly in other developing countries

with similar digital access patterns, could offer valuable insights into how local norms influence both cyberbullying and the reception of digital monitoring tools.

Third, there is an opportunity to expand the scope of AI tools used in cyberbullying detection. While this study relied primarily on text-based analysis, future systems could incorporate multimodal analytics, including voice, image, and video content, especially given the increasing prevalence of harassment through memes, doctored images, and video manipulation. Research into ethical and privacy-sensitive deployment of such multimodal systems would be a worthwhile direction, particularly when deployed in educational environments.

Another area for exploration is the personalization of awareness content. Adaptive learning systems powered by machine learning could tailor digital citizenship lessons to individual students based on their online behavior patterns and comprehension levels. Such systems would require careful design to ensure that personalization does not become invasive or stigmatizing, but they offer the potential to make awareness training more relevant and engaging.

Additionally, future studies should engage student voices more directly in the design and governance of online safety tools. Participatory research models, where students co-create intervention strategies or contribute to ethical oversight committees, could increase buy-in and foster a deeper sense of responsibility for maintaining safe digital spaces. This aligns with the growing emphasis on youth-led initiatives in the global digital rights movement (Livingstone, 2021).

Finally, there is a clear need for more robust policy evaluations that examine how national cyber laws, school-level digital conduct policies, and international e-safety frameworks interact with technological solutions. Investigating the alignment—or lack thereof—between legal protections and everyday digital experiences of youth could inform more effective and equitable policy design.

In essence, the future of cyberbullying prevention lies not in any one solution but in a dynamic, interdisciplinary ecosystem that bridges educational practice, technological innovation, ethical governance, and youth empowerment. With sustained investment and collaboration across sectors, it is possible to build digital environments where safety, respect, and resilience are not only encouraged but expected.

Acknowledgment

The authors would like to express their sincere gratitude to the school administrators, teachers, students, and parents who participated in this study and generously contributed their time, insights, and experiences.

Special thanks are extended to the principals and digital safety coordinators of the six participating schools for their cooperation and trust during the data collection and intervention phases.

Disclosure of Interest

The authors declare that there is no conflict of interest related to the publication of this research. No commercial, financial, or personal relationships influenced the design, implementation, analysis, or reporting of the findings presented in this study.

Funding Information

This research was carried out without any financial support from funding agencies, institutions, or commercial organizations. The authors confirm that the study was conducted using personal or institutional resources, and no specific grant or project funding was received from public, private, or non-profit sectors during this research and its publication process.

References

Aboujaoude, E., Savage, M. W., Starcevic, V., & Salame, W. O. (2015). Cyberbullying: Review of an old problem gone viral. Journal of Adolescent Health, 57(1), 10–18. https://doi.org/10.1016/j.jadohealth.2015.04.011

Bandura, A. (1977). Social learning theory. Englewood Cliffs, NJ: Prentice-Hall.

Barlett, C. P., Sykes, T. N., & Sengupta, A. (2021). Investigating the effectiveness of a brief, school-based cyberbullying prevention program. Journal of School Violence, 20(1), 21–33. https://doi.org/10.1080/15388220.2020.1848953

Bauman, S. (2013). Cyberbullying: What counselors need to know. Alexandria, VA: American Counseling Association.

Bauman, S., Cross, D., & Walker, J. (2013). Principles of cyberbullying research: Definitions, measures, and methodology. Routledge.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. Spine, 25(24), 3186–3191.

Betts, L. R., & Spencer, K. A. (2017). "People think it's a harmless joke": Young people's understandings of the impact of technology, digital vulnerability, and cyberbullying in the UK. Children & Society, 31(2), 112–123. https://doi.org/10.1111/chso.12189

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chen, J. K., Ho, S. S., & Liao, Y. (2017). A longitudinal study of cyberbullying perpetration: The role of interpersonal and intrapersonal factors. School Psychology International, 38(5), 586–605. https://doi.org/10.1177/0143034317727745

Chisholm, J. F. (2015). Review of the status of cyberbullying and cyberbullying prevention. Journal of Information Systems Education, 26(2), 95–102.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Dredge, R., Gleeson, J., & de la Piedad Garcia, X. (2014). Cyberbullying in social networking sites: An adolescent victim's perspective. Computers in Human Behavior, 36, 13–20. https://doi.org/10.1016/j.chb.2014.03.026

Durlak, J. A., Domitrovich, C. E., Weissberg, R. P., & Gullotta, T. P. (Eds.). (2015). Handbook of social and emotional learning: Research and practice. New York, NY: Guilford Press.

Fahy, A. E., Stansfeld, S. A., Smuk, M., Smith, N. R., & Green, J. (2016). Longitudinal associations between cyberbullying involvement and adolescent mental health. Journal of Adolescent Health, 59(5), 502–509. https://doi.org/10.1016/j.jadohealth.2016.06.006

Ferreira, W., & Waseem, Z. (2022). Challenges in developing fair and accurate cyberbullying detection systems. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Dublin, Ireland.

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—principles and practices. Health Services Research, 48(6pt2), 2134–2156. https://doi.org/10.1111/1475-6773.12117

Floridi, L., & Cowls, J. (2022). The ethics of artificial intelligence: Key concepts and debates. Oxford University Press.

Greene, J. C. (2007). Mixed methods in social inquiry. San Francisco, CA: Jossey-Bass.

Hamm, M. P., Newton, A. S., Chisholm, A., Shulhan, J., Milne, A., Sundar, P., ... & Hartling, L. (2015). Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies. JAMA Pediatrics, 169(8), 770–777. https://doi.org/10.1001/jamapediatrics.2015.0944

Hinduja, S., & Patchin, J. W. (2019). Connecting adolescent suicide to the severity of bullying and cyberbullying. Journal of School Violence, 18(3), 333–346. https://doi.org/10.1080/15388220.2018.1492417

Jiang, M., Sun, L., & Huang, Z. (2020). Challenges of automated cyberbullying detection: Toward robust and inclusive NLP models. ACM Transactions on the Web, 14(4), 1–25.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. Psychological Bulletin, 140(4), 1073–1137. https://doi.org/10.1037/a0035618

Kowalski, R. M., Limber, S. P., & McCord, A. (2021). A developmental approach to cyberbullying: A review of defining characteristics and a proposal for a socio-ecological model. Journal of Adolescent Health, 68(6), 1061–1070. https://doi.org/10.1016/j.jadohealth.2020.12.018

Livingstone, S., & Stoilova, M. (2021). The digital environment and children's well-being: A conceptual framework. UNICEF Office of Research–Innocenti.

O'Brien, T., & DeLongis, A. (2022). Cyberbullying and the adolescent brain: Investigating psychological mechanisms of digital aggression. Child and Adolescent Psychiatry and Mental Health, 16(1), 1–9.

Patchin, J. W., & Hinduja, S. (2018). Sexting as an emerging concern for adolescent health: A review of the literature. Pediatrics, 141(6), e20173126. https://doi.org/10.1542/peds.2017-3126

Sarkar, K. (2022). Multilingual cyberbullying detection using transformer-based deep learning models. International Journal of Information Security, 21(4), 455–468.

Shariff, S., & Gouin, R. (2015). Cyberbullying prevention: Youth perspectives on the role of parents, educators, and the law. Canadian Journal of Education, 38(3), 1–32.

Slonje, R., Smith, P. K., & Frisén, A. (2013). The nature of cyberbullying, and strategies for prevention. Computers in Human Behavior, 29(1), 26–32. https://doi.org/10.1016/j.chb.2012.05.024

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in Human Behavior, 26(3), 277–287. https://doi.org/10.1016/j.chb.2009.11.014

UNESCO. (2022). Guidelines for teachers on online safety and digital well-being. Paris, France: United Nations Educational, Scientific and Cultural Organization.

Williford, A., Boulton, A. J., & Noland, B. L. (2013). The role of school climate in combating cyberbullying. Theory Into Practice, 52(4), 297–302. https://doi.org/10.1080/00405841.2013.829735

Wright, M. F. (2018). Parental mediation, cyberbullying, and cyber victimization: A longitudinal analysis. Child & Youth Care Forum, 47(5), 693–709. https://doi.org/10.1007/s10566-018-9455-6

Yang, K., Yao, Y., & Tannenbaum, A. (2018). Ethical and privacy concerns in artificial intelligence monitoring in education. AI & Society, 34(3), 569–575.

Appendix

Description of the Digital Citizenship Curriculum

The digital citizenship curriculum implemented in the two intervention schools was structured as a six-week program, with one 45-minute session conducted each week. The content was adapted to suit local cultural and linguistic contexts while aligning with internationally recognized frameworks such as Common Sense Education and UNESCO's digital citizenship guidelines. Below is a summary of the curriculum topics:

Week 1: Introduction to Cyberbullying

Students were introduced to the concept of cyberbullying, its forms (e.g., flaming, outing, exclusion), and its impact on individuals. Local case studies were used to contextualize discussions.

Week 2: Empathy and Online Communication

Activities included role-playing exercises and reflection activities designed to help students understand how their online words and actions can affect others.

Week 3: Privacy, Digital Footprints, and Consent

Students learned about managing personal information, understanding digital footprints, and the importance of obtaining consent before sharing content.

Week 4: Bystander Intervention and Peer Support

Focus was placed on encouraging students to speak up, report harmful behavior, and support peers in distress, including how to use the anonymous peer-reporting tool.

Week 5: Cyber Laws and Reporting Mechanisms

Students were introduced to Pakistan's Prevention of Electronic Crimes Act (PECA, 2016), and were informed about how and where to report cyberbullying legally and within their schools.

Week 6: Responsible Online Behavior and Digital Ethics

This session emphasized ethical decision-making, the long-term consequences of online behavior, and the importance of mutual respect and inclusivity in digital spaces.

SafeTalk AI Tool Technical Configuration

The SafeTalk platform was customized with the following configurations for this study:

- **Language Models:** Fine-tuned multilingual BERT variant trained on a hybrid corpus of English, Urdu, and transliterated Roman Urdu.
- **Detection Threshold:** Set at a toxicity probability score of ≥0.70 to balance sensitivity and false-positive reduction.
- **Anonymization Protocol:** All student identifiers were hashed before being processed by the platform; raw data remained encrypted and accessible only to authorized safeguarding staff
- **Dashboard Access:** Restricted to school counselors and designated teachers, with login.
- **User Prompts:** Custom messages such as "Please reconsider your message" or "Is this respectful?" were shown when potential harm was detected pre-transmission.

Sample Survey Items

"Have you ever been called hurtful names or insulted via text or social media in the past three months?"

"How often do you witness peers being bullied online?"

"I feel confident reporting cyberbullying incidents anonymously at school."

"I understand the consequences of sharing personal information online."

Responses were recorded on a 5-point Likert scale (1 = Never, 5 = Very Often).

Open Access Statement