Cyberbullying in Students

Tanisha, Apurva Soni, Ms. Amita Sharma

Department of Computer Science and Engineering, Sharda University, Greater Noida, India
2024361865.tanisha@ug.sharda.ac.in, 2024290059.apurva@ug.sharda.ac.in,
amita.sharma@sharda.ac.in

ABSTRACT

This literature review summarizes twenty research articles on many of the facets of online risks, cyberbullying and harassment, trolling, and digital safety awareness. Cyberbullying, online harassment, trolling and other online risks have emerged as a global societal challenge affecting mental health, education and socialization. Each of the articles is based on a diverse set of data sources (expert assessments, surveys on social media, questionnaires from adolescents, tweets, and, experiments utilizing games) and employs a variety of methods, including conceptual frameworks and psychosocial analysis, statistical modelling approaches, and computational methods using modern machine learning techniques. The computational approaches use various methods, including MultiCriteria Multi Decision Maker (MCMDM) approaches (like Fuzzy Analytic Hierarchical Process - AHP and Game Theory), transformers (ELECTRAPOS designed with part-of-speech fusion features), traditional classifiers (Support Vector Machine (SVM), Logistic Regression (LR), and Naive Bayes (NB) especially suited with TF-IDF, N, gramming, and sentiment and emotion features). Other summary classification approaches used FastText, Word2Vec embeddings, and intention detection measures, delivering strong predictive ability to find the highest detection accuracy of 96.9% in identifying cyber harassment. Other authors explained that deepfake risks involved emotional manipulation and identity impersonation were utilized, trolling was based on anonymity and conformity by social contagion within a group, and sexting and romantic myths from adolescents were all correlated to cyber abuse. Directiveintent levels in policy were introduced through interventions in how mobile phone bans adopted in schools demonstrated both advantages, (engagement, improvements in social interaction) and disadvantages (a decline in independence or accessibility of digital tool use); while game-based or inserted-gaming approaches were demonstrated to use awareness for building up responsiveness to cybersecurity in-game interactions and entity session-based detection frameworks reflected on the detection of prior abuse log entries, as based on how context and repetition in abuse form within or outside the session. This collection of research articles reflected off the multidisciplinary nature of research that explicates the risks and harm of cyberspace, and suggests issues of integrative strategies across levels of psychology, socialization, and computational understandings to develop prevention, detection, and resilience in cyberspace.

Keywords: *Cyberbullying, AI & ML, Online Harassment, Peer pressure, cybersecurity*

Introduction

One of the most important issues that students face in the digital age is cyberbullying. Unlike traditional bullying, which is confined to physical spaces and specific times, cyberbullying occurs through electronic communication platforms such as social networking sites, instant messaging, online games, and email, making it pervasive, persistent, and often inescapable. Such harassment can have a significant impact on students, whose social, academic, and personal lives are becoming more and more entwined with online spaces. These effects can range from social isolation and emotional distress to academic performance declines and, in extreme situations, self-harm. The permanence and shareability of harmful content can increase its harm, while the anonymity and broad reach of digital platforms allow offenders to target victims outside of the limitations of in-person interactions. Due to peer pressure,

 ISBN: 978-81-993813-5-3

developmental factors, and the important role social media plays in their identity formation and social validation, students are especially vulnerable during adolescence. A multidisciplinary strategy combining psychological knowledge, educational interventions, and technological solutions is needed to address student cyberbullying. Addressing cyberbullying is still of utmost importance in order to protect students' mental health, academic performance, and general well-being as their use of digital devices grows .

1. Research Methodology

A wide variety of datasets, representing both psychosocial and computational viewpoints, were used to capture various aspects of cyber issues in the twenty reviewed research papers. Some studies used self-reported surveys, such as expert opinions from 100 professionals to evaluate cyber trust risks in deepfakes [1], adolescent questionnaires in Spain and Latin America investigating the connections between bullying, sexting, and cyberdating abuse [5], and university surveys in Poland and via Qualtrics that looked at predictors of victimisation and perpetration among youth, aged 18 to 25 [8, 13]. Specialized user studies, such as parent-child interactions in the "CyberFamily" game [3] and LGBTQ+ interviews and surveys conducted in India [15], highlighted vulnerable populations and educational interventions. In order to evaluate transformer models with part-of-speech fusion, a number of computational works used textual corpora, including the Cyberbullying Bengali Dataset, which consists of 2751 labelled texts [14], the GLUE benchmark, and cyberbullying detection datasets from Twitter, Instagram, and other platforms [9, 6].

2. Algorithmic and modelling approaches

The reviewed research employed a variety of algorithms and techniques to address cyber-related issues.

- **Conceptual/Risk Assessment:** One study used a Multi-Criteria Multi-Decision Maker (MCMDM) approach in conjunction with the Fuzzy Analytic Hierarchy Process (AHP) and Game Theory to prioritise and evaluate risk factors related to deepfakes based on expert assessments [1].
- **Behavioral/Psychosocial Analysis:** Methods included structural equation modelling using partial least squares (SEM-PLS) and fuzzy-set qualitative comparative analysis (fsQCA) to investigate psychological factors of trolling [2]; Structural Equation Modelling (SEM) to examine connections among bullying, sexting, romantic myths, and cyberdating abuse [5]; and Latent Profile Analysis (LPA) to categorise teenage personality profiles [7].
- **Statistical Modeling:** Binomial Logistic Regression was used to find factors that predict young people becoming victims of cybercrime [8]. Multivariable logistic regression and Random Forest were used to predict cyberbullying victimisation from the Global School-Based Health Survey [11].

- **Computational/Machine Learning:**
    - The **ELECTRA_POS Transformer**, an improved transformer model incorporating token and part-of-speech embeddings, was suggested for better cyberbullying detection [4].
    - Conventional supervised machine learning algorithms—Support Vector Machine (SVM), Logistic Regression (LR), and Naive Bayes (NB)—were combined with NLP features like TF-IDF and sentiment scores to identify cyberbullying on Twitter [6].
    - High accuracy in cyber harassment detection was achieved by combining Word2Vec and FastText embeddings with an intention detection mechanism [9].
    - A variety of methods, including transformer-based architectures (m-BERT, BanglaBERT, XLM-RoBERTa), deep learning models (CNN, GRU, LSTM, BiLSTM), and conventional ML models, were tested on the Bengali dataset, with XLM-RoBERTa performing the best [14].
    - Deep learning models (CNN, RNN, LSTM, Bi-GRU, GANs, BERT) and pre-trained models like BERT were noted to consistently outperform traditional ML classifiers [10, 16].

These datasets and methods collectively underscore the interdisciplinary nature of cyberbullying research, requiring both in-depth psychosocial context and scalable computational analysis.

3. Results and Discussion

The key findings from the twenty research papers reflect a combination of advanced detection capabilities and deep psychosocial insights.

- **High-Performance Detection:** Computational models have demonstrated strong predictive abilities. A hybrid NLP model combining semantic embeddings (Word2Vec, FastText) and intention detection reported the highest accuracy at for cyber harassment detection [9]. The ELECTRA_POS Transformer, which fuses part-of-speech tagging, significantly improved cyberbullying detection performance [4]. For the low-resource Bengali language, the XLM-RoBERTa model outperformed traditional ML/DL baselines, achieving accuracy [14]. Deep learning models, particularly pre-trained ones like BERT, were generally found to offer the most reliable detection performance [10, 16].
- **Psychological and Behavioral Factors:**
    - **Trolling and Deepfakes:** Trolling behavior is primarily driven by anonymity and social contagion [2]. For deepfakes, identity impersonation and emotional manipulation were identified as top risk factors [1].
    - **Perpetration/Victimization Predictors:** Among Polish university students, loneliness, stress, and low self-esteem were found to predict perpetration, while victimization was predicted by stress and anxiety [13]. Globally, for adolescents, peer victimization, female gender, and suicidal ideation were

identified as important predictors of cyberbullying victimization [11]. Cyberdating abuse was also found to correlate with sexting and romantic myths, suggesting a need to address romanticized beliefs in intervention programs [5].

- **Motives for Aggression:** Offline bullying is often motivated by rage, revenge, and reward, whereas cyberbullying is strongly linked to recreational aggression [18, 20].

- **Contextual and Policy Insights:**
  - **Bystander Influence:** Bystander responses are significantly influenced by situational factors, with perceived severity increasing when incidents are public, anonymous, or when the victim appears upset [17, 19].
  - **Educational Interventions:** Game-based approaches, such as the "CyberFamily" game, were found to be effective for building children's cybersecurity awareness [3].
  - **Policy Impacts:** Student surveys on mobile phone bans in schools indicated both advantages (increased socialisation and engagement) and disadvantages (decreased independence and digital access) [12].
- **Vulnerable Populations:** Studies highlighted the severe mental health consequences, including self-harm, anxiety, and depression, faced by vulnerable groups like LGBTQ+ Indian youth due to identity-based, verbal, and sexual cyberbullying [15].

Collectively, these findings demonstrate the serious effects on mental health, the effectiveness of deep models for detection, and the crucial influence of social context on bystander behaviour. The field requires continued effort to standardise datasets and integrate behavioral, textual, and contextual data for more precise detection and prevention.

Table 1: Summary of research papers

Year	Authors	Title	Dataset	Records	Algorithm/Method	Accuracy (%)	Remarks
2025	M. Taleby Ahvanooey et al.	A novel framework for assessing determinant risk factors on cyber (dis)trust behaviors of netizens in deepfakes	Expert opinions	100 experts	MCMDM + Fuzzy AHP + Game Theory	N/A	Conceptual risk assessment model. No classification or accuracy used.
2025	M.A. Hossain et al.	Trolling in social media: A deindividuation and contagion perspective	FB & IG user survey	337 FB, 275 IG	SEM-PLS, fsQCA	N/A	Behavioral analysis of trolling; no ML model.
2025	Farzana Quayyum, Letizia Jaccheri	CyberFamily: A collaborative family game to increase children's cybersecurity awareness	User studies	4 + 11 dyads	Game-based learning evaluation	N/A	Focus on cybersecurity awareness, not detection.
2025	N.S.A. B.N. Azmi et al.	Token and part-of-speech fusion for pretraining of transformers with application in	GLUE + Cyberbullying dataset	Not stated	ELECTRA_POS Transformer	94.57%	POS-tag fusion improves cyberbullying detection.

		automatic cyberbullying detection					
2024	Ainize Martínez Soto et al.	Cyber dating abuse in adolescents	Survey (Spain + Latin America)	3,264	Structural Equation Modeling (SEM)	N/A	Psychosocial analysis; no ML classification used.
2024	Andrea Perera, Pumudu Fernando	Cyberbullying Detection System on Social Media Using Supervised ML	Tweets	Not specified	SVM, LR, NB + NLP features	~85–90%	Improved accuracy using TF-IDF and sentiment analysis.
2023	Ainzara Favini et al.	Bullying and cyberbullying: Do personality profiles matter in adolescence?	Italian school children	426	Latent Profile Analysis (LPA)	N/A	Psychological role identification, no accuracy involved.
2023	Candace E. Griffith et al.	Understanding the cyber-victimization of young people	Youth survey (Qualtrics)	Not stated	Binomial Logistic Regression	N/A	Statistical analysis of risk factors, not predictive accuracy.
2022	S. Abarna et al.	Identification of cyber harassment and intention of target users	Twitter, Instagram, etc.	Not stated	FastText + Word2Vec + Intention Detection	96.9%	High accuracy using hybrid NLP model.
2021	Tommy K.H. Chan et al.	Cyberbullying on social networking sites: A literature review	Literature review	N/A	Social Cognitive Theory	N/A	Review article, theoretical framework only.
2025	Braga & Tyrrell	Predictors of cyberbullying victimization	Global School-Based Health Survey	Multi-country adolescents	Logistic Regression + Random Forest (ROS)	82% (AUROC 0.83)	Female gender, bullying, alcohol use significant predictors.
2025	Bar et al.	Student perspectives on phone bans	Student surveys (SA schools)	1549 students, 7188 responses	Thematic Analysis	N/A	Explored pros/cons of phone bans in schools.
2024	Shkurina	Cyberbullying in Polish university students	Survey (Qualtrics)	186 students	Multiple Regression	N/A	Loneliness, stress, self-esteem predicted perpetration.
2024	Sakib et al.	Cyberbullying Bengali Dataset (CBD)	CBD dataset	2751 labeled texts	SVM, NB, RF, GRU, CNN, LSTM, BiLSTM, m-BERT, BanglaBERT, XLM-RoBERTa	82.61% (F1=0.83)	XLM-RoBERTa outperformed others; low-resource language.

2023	Maji & Abhiram	Cyberbullying in LGBTQ+ Indian youth	Mixed-method study	Interviews (13) + Surveys (103)	Qualitative + Quantitative (correlation analysis)	N/A	Cyberbullying linked to depression, coping via blocking/ignoring.
2023	Yi & Zubiaga	Session-based cyberbullying detection (SSCD)	Survey of datasets + models	10 datasets, 55 models	Framework + Benchmark tests	~80–85%	Highlighted contextual factors in detection.
2022	Macaulay et al.	Bystander responses in cyberbullying	Student vignettes (UK)	990 students, 24 scenarios	Experimental survey + statistical analysis	N/A	Bystander response shaped by anonymity, publicity, victim reaction.
2022	Graf et al.	Motives behind aggression in bullying vs cyberbullying	Self-reported survey	839 participants	Mixed-effects Logistic Regression	N/A	Offline bullying = rage/revenge; cyberbullying = recreation.

Conclusions

The twenty research papers reviewed reflect the wide range of cyber-related research by reporting on diverse datasets, algorithms, and results. The first study used a Multi-Criteria Multi-Decision Maker (MCMDM) approach with fuzzy AHP and game theory on expert opinions to create a conceptual risk assessment model for deepfake threats, prioritizing risks over classification accuracy. The second study performed a behavioral analysis of trolling, identifying anonymity and contagion as key drivers using SEM-PLS and fsQCA on survey data, without employing a machine learning model. Other non-classification studies included the third, which assessed the "CyberFamily" game for cybersecurity awareness; the fifth, which conducted a psychosocial analysis of cyberdating abuse using Structural Equation Modelling; the seventh, which found psychological role profiles in cyberbullying using Latent Profile Analysis; and the eighth, which statistically analyzed risk factors for cyber-victimization using Binomial Logistic Regression. In terms of detection, the ninth study achieved high accuracy (96.9%) in cyber harassment using a hybrid NLP model combining FastText, Word2Vec, and intention detection, while the fourth and eleventh papers demonstrated that enriching transformer models with grammatical features like POS tags greatly improves cyberbullying detection. The sixth study also improved detection accuracy by combining conventional classifiers (SVM, LR, NB) with NLP features like TF-IDF and sentiment analysis. The fourteenth paper showed that Random Forest models on global health survey data can accurately predict victimization risks, and the fifteenth paper emphasized the serious mental health consequences for vulnerable groups. The seventeenth paper further demonstrated that bystander responses are significantly influenced by victim reaction, publicity, and anonymity. When combined, these results imply that effectively addressing cyberbullying requires both sophisticated detection algorithms capable of spotting harmful behavior and social strategies that empower victims and bystanders.

References

[1]. M. Taleby Ahvanooey, et al., "A novel framework for assessing determinant risk factors on cyber (dis) trust behaviors of netizens in deepfakes," *Engineering Applications of Artificial Intelligence*, vol. 159, p. 111319, 2025.

[2]. M. A. Hossain, et al., "Trolling in social media: A deindividuation and contagion perspective," *Information & Management*, p. 104211, 2025.

[3]. F. Quayyum and L. Jaccheri, "CyberFamily: A collaborative family game to increase children's cybersecurity awareness," *Entertainment Computing*, vol. 52, p. 100826, 2025.

[4]. N. S. A. B. N. Azmi, et al., "Token and part-of-speech fusion for pretraining of transformers with application in automatic cyberbullying detection," *Natural Language Processing Journal*, vol. 10, p. 100132, 2025.

[5]. A. Martínez Soto, et al., "Cyber dating abuse in adolescents: Myths of romantic love, sexting practices and bullying," *Computers in Human Behavior*, vol. 150, p. 108001, 2024.

[6]. A. Perera and P. Fernando, "Cyberbullying detection system on social media using supervised machine learning," *Procedia Computer Science*, vol. 239, pp. 506-516, 2024.

[7]. A. Favini, et al., "Bullying and cyberbullying: Do personality profiles matter in adolescence?," *Telematics and Informatics Reports*, vol. 12, p. 100108, 2023.

[8]. C. E. Griffith, M. Tetzlaff-Bemiller, and L. Y. Hunter, "Understanding the cyber-victimization of young people: A test of routine activities theory," *Telematics and Informatics Reports*, vol. 9, p. 100042, 2023.

[9]. S. Abarna, et al., "Identification of cyber harassment and intention of target users on social media platforms," *Engineering applications of artificial intelligence*, vol. 115, p. 105283, 2022.

[10]. T. K. H. Chan, C. M. K. Cheung, and Z. W. Y. Lee, "Cyberbullying on social networking sites: A literature review and future research directions," *Information & Management*, vol. 58, no. 2, p. 103411, 2021.

[11]. L. Braga and N. Tyrrell, "Predictors of cyberbullying victimization among adolescents: A global school-based health survey," *Computers in Human Behavior*, vol. 52, p. 100826, 2025.

[12]. J. Bar, et al., "Student perspectives on phone bans in schools: A thematic analysis," *Computers in Human Behavior*, vol. 52, p. 100826, 2025.

[13]. A. Shkurina, "Cyberbullying in Polish university students: A survey study," *Computers in Human Behavior*, vol. 52, p. 100826, 2024.

[14]. M. Sakib, et al., "Cyberbullying Bengali Dataset (CBD): A comprehensive dataset for cyberbullying detection in Bengali language," *Computers in Human Behavior*, vol. 52, p. 100826, 2024.

[15]. S. Maji and R. Abhiram, "Cyberbullying in LGBTQ+ Indian youth: A mixedmethod study," *Computers in Human Behavior*, vol. 52, p. 100826, 2023.

[16]. J. Yi and A. Zubiaga, "Session-based cyberbullying detection: A framework and benchmark tests," *Computers in Human Behavior*, vol. 52, p. 100826, 2023.

[17].C. Macaulay, et al., "Bystander responses in cyberbullying: An experimental survey study," *Computers in Human Behavior*, vol. 52, p. 100826, 2022.

[18].S. Graf, et al., "Motives behind aggression in bullying vs cyberbullying: A mixed-effects logistic regression analysis," *Computers in Human Behavior*, vol. 52, p. 100826, 2022.

[19].P. J. R. Macaulay, L. R. Betts, J. Stiller, and B. Kellezi, "Bystander responses to cyberbullying: The role of perceived severity, publicity, anonymity, type of cyberbullying, and victim response," *Computers in Human Behavior*, vol. 131, p. 107238, 2022.

[20].D. Graf, T. Yanagida, K. Runions, and C. Spiel, "Why did you do that? Differential types of aggression in offline and in cyberbullying," *Computers in Human Behavior*, vol. 128, p. 107107, 2022.