Towards Safer Social Media: A Survey on Artificial Intelligence for Multimodal Cyber Bullying Detection

Premalatha G.
Department of Information Technology,
Puducherry Technological University, Puducherry

G. Santhi
Department of Information Technology,
Puducherry Technological University, Puducherry

ABSTRACT
The growth of social media has led to an increase in the scale of cyber harassment, from texts to images, videos, memes, deepfakes, and identity-based assaults. Traditional machine learning approaches are limited to identify such multimodal and context-dependent misuse. This study gives a complete assessment of AI based cyber harassment detection including datasets, preprocessing, deep learning, transformer models, multimodal fusion, big language models, explainable AI, federated learning, and privacy-preserving frameworks. Transformer based and multi-modal techniques offer better contextual and semantic comprehension compared to conventional approaches. The poll also highlights important research problems, such as multilingual detection, bias reduction, explainability, real-time scalability, and the identification of developing AI-generated offensive material.

Keywords
Detection of Cyberbullying, Social Media Harassment, Multimodal Deep Learning, Large Language Models (LLMs), Artificial Intelligence.

1. INTRODUCTION
The fast expansion of social media has changed the way people connect, exchange information, study and do business. [1,2] Platforms like Facebook, Instagram, X, YouTube, TikTok, and WhatsApp have become essential that affect public opinion, education, health awareness, and economic activity throughout the globe. However, the fast development has also raised severe issues over security and privacy [3,4], in spite of the many advantages it provides.

Among the many hazards posed by social media, cyberbullying has emerged as one of the most common types of online abuse **(Figure 1)** [5]. Cyberbullying is different from traditional bullying since it is not limited by time and place, and damaging material may spread continually and to a broader audience [6]. A significant challenge is the recognition of cyberbullying on social media platforms [7, 8]. Existing moderation frameworks are not enough to address complexity and variety. The constraints have led to the introduction of Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) based systems that can automatically discover patterns from large scale data.

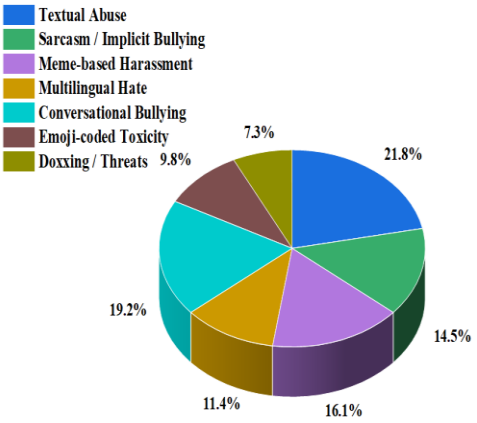

Figure 1. Cyberbullying Issue Categories Bar (2021-2025).

Artificial intellect is a computational system that mimics human intellect and accomplishes tasks. Machine Learning allows computers to understand patterns from data and make better predictions without being explicitly programmed [9]. Traditional machine learning approaches have been extensively employed for cyberbullying detection, yet they typically fail to capture complex context [10]. To circumvent these restrictions, Deep Learning (DL) models have been proposed since they can learn hierarchical features automatically from raw data [11-17]. Despite the remarkable achievements in the performance of cyberbullying detection by ML and DL techniques, several difficulties are still open [10]. Explainability problems, privacy preservation, and ethical deployment are still the obstacles to the actual application of intelligent detection systems [16,17]. Thus, the use of powerful Machine Learning and Deep Learning techniques is an important research topic for the enhancement of online safety. This review presents an in-depth analysis of the recent advances, obstacles, and prospects for AI-based social media harassment and cyberbullying detection systems published in 2021-2026 with particular focus on machine learning, deep learning, multimodal and adaptive detection frameworks.

2. RELATED WORK
2.1 Types of Social Media Harassment
Social media harassment encompasses hate speech, trolling, cyberbullying, cyberstalking, sexual harassment, doxing, body shaming, and political and identity-based abuse. Advances in AI and deep learning have resulted in online harassment becoming multimodal, ranging from text, photos, videos, audio, memes to emojis. This has increased the demand for effective automated AI based harassment detection systems.

For example, **F. F. Nova et al. [18]** examined online abuse of disadvantaged Hijra people in Bangladesh. **E. A. Vogels et al. [19]** found that about half of all teens in the United States had experienced the appearance-based harass. **S. Schoenebeck et al. [20]** studied several types of online harassment and their

perceived harms on social media. **G. Ray et al. [21]** did a comprehensive assessment of the existing literature on cyberbullying and noted that cyberbullying is still one of the most widespread kinds of online harassment among teenagers and young adults. **R. Saha et al. [22]** evaluated the progress of the research on online abuse, and depicted online abuse as a complex phenomenon that includes damaging digital interactions, cyber violence and abusive online behaviors. **J. Joseph et al. [23]** studied violence against women in digital spaces and identified developing kinds of online harassment such as cyberstalking and technology enabled abuse. **T. M. Milyane [24]** studied the adolescent's perception of cyberbullying on social media platforms. The results revealed that negative remarks and insults were the most prevalent kinds of cyberbullying encountered by teenagers. The study by **I. Bachmann et al. [25]** showed LGBTQ+ people and marginalized groups are disproportionately targeted via gendered harassment, discrimination and online violence. **P. Pérez et al. [26]** studied Online Violence Against Women (OVAW) and listed a number of growing types of online abuse such as cyberstalking, sextortion. **P. Yi et al. [27]** underline the significance of considering cyberbullying as a sequence of social interactions, not only individual offending texts, and the need of contextual modeling and conversation dynamics in detection systems. **G. Akwa et al. [28]** further observed that the rapid adoption of transformer architectures, multi-modal learning and large language models has contributed to significantly better automated cyberbullying detection, while also presenting new challenges in the domains of explainability, fairness, privacy preservation and real-time deployment.

Cyberbullying is the most researched online harassment kind, leading to the development of detection systems based on classical machine learning, deep learning, transformer, multimodal and large language model (LLM) approaches. A rigorous characterisation of these strategies is necessary to construct scalable, adaptive and context-aware moderation systems.

2.2 Hierarchical Classification of Social Media Harassment Detection Frameworks

Cyber harassment detection systems can be generally categorized into three main classes **(Figure 2): (i) Text-Based Attack Detection**, which deals with abusive language in written form such as comments, messages, hate speech and cyberbullying; **(ii) Image-Based Attack Detection**, which identifies malicious content in images, memes, screenshots and visual posts; **(iii) Video-Based Attack Detection,** which tackles harassment via multimedia and audio-visual content. Understanding these detecting modalities is necessary to design comprehensive cyber harassment mitigation systems capable of managing changing online abusive activities.

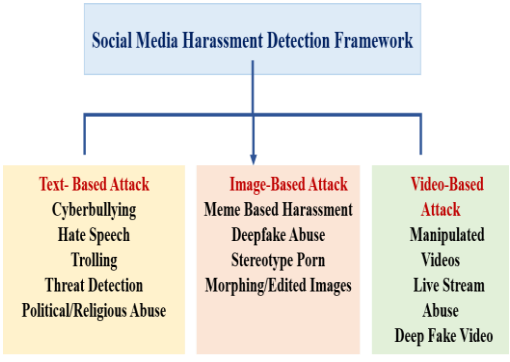

Figure 2. Social Media Harassment Detection Framework.

2.2.1 Text-based Attack Detection

Text-based attack detection is one of the most explored fields. The growing usage of informal language, code-mixed text and context-dependent phrases has made the automated identification substantially complex. Recent developments in Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL), transformer architectures such as BERT and RoBERTa, and Large Language Models (LLMs) have significantly enhanced the ability to detect subtle and emerging online abuse patterns **(Figure 3).**

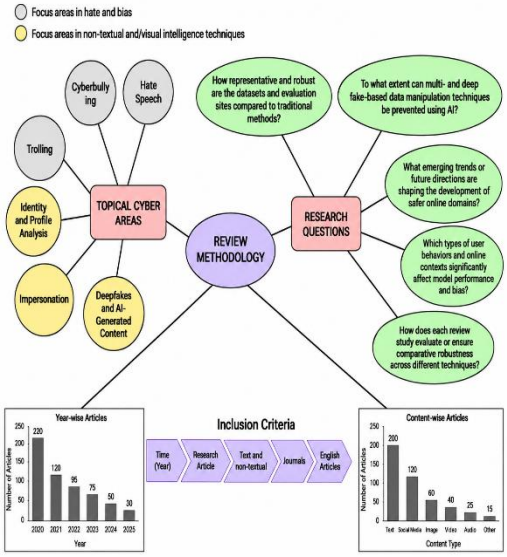

Figure 3. Conceptual overview of categories of textual cyber abuse (Reproduced and adapted with permission from J. A. Diaz-Garcia et al. [29].

A. Johri et al. [30] noted that social media abuse comes in several forms. The research pointed to the increased complexity and incidence of internet harms. Similarly**, A. Alabdulwahab et al. [31]** pointed out the limits of standard content moderation methods in recognizing sarcasm, slang, and context-dependent toxicity, arguing that deep learning-based models provide better options for spotting nuanced harmful behaviors. **C. P. Barlett et al. [32]** proposed the Integrative Cyberbullying Theory (ICBT) as an overarching paradigm to explain the relationship between cyberbullying perpetration and victimization and psychological outcomes. **S. Sree et al. [33]** proposed a deep learning system BullyNet. They found that modern cyber harassment is ever-changing in terms of shifting language patterns, subtle aggressiveness and context sensitive toxicity. Cyberbullying was studied by **J. Pyzalski et al. [34]** using qualitative interviews with teenagers aged 13–16 years. The authors stressed the need for preventative methods aimed at developing empathy, increasing bystander involvement, and strengthening teacher-student connections to decrease the occurrence of cyberbullying. **S. Kim et al. [35]** presented a detailed survey of machine learning-based cyberbullying detection. They studied how to include aspects of human behaviour into algorithmic detection schemes and examined their alignment with current theories of cyberbullying. **F. Elsafoury et al. [36]** conducted a thorough study of automated cyberbullying detection strategies including the whole machine learning pipeline, from data collection to model evaluation. **T. Mahmud et al. [37]** studied cyberbullying detection for multilingual and low-resource language environments, with a focus on Bangla and Chittagonian languages. The authors constructed a dedicated sarcasm dataset and integrated explainable artificial intelligence (XAI) utilizing

Local Interpretable Model-Agnostic Explanations (LIME) to improve transparency and detection capabilities. Likewise, **A. Abdullah et al. [38]** emphasized that transformer-based detection strategy has shown better performance in identifying nuanced, context-dependent and complicated harassment categories in comparison to traditional machine learning approaches. **A. B. A. H. Murshed et al. [39]** proposed a hybrid deep learning framework called DEA-RNN, which combines Elman Recurrent Neural Networks and an improved Dolphin Echolocation Algorithm to improve the identification of cyberbullying on Twitter. The model they suggested performed better than typical machine learning and deep learning methods, with an accuracy of 90.45%. **E.Y. Daraghmi et al. [40]** tried to resolve the deficit of non-English languages in cyberbullying research, by suggesting an integrated framework of CNN-BiLSTM-GRU for Arabic cyberbullying detection. **S. Sihab-Us-Sakib et al. [41]** have further expanded multilingual cyberbullying detection with an emphasis on low-resource languages such as Bengali with XLM-RoBERTa achieved the best performance, indicating the usefulness of the transformer architecture for multilingual and context-aware cyberbullying detection. A mixed machine learning approach for Bengali cyberbullying detection was suggested by **A. Akhter et al. [42].** The work included a thorough preprocessing stage, TF-IDF feature extraction and Instance Hardness Threshold (IHT) resampling for handling class imbalance prior to the classification stage. Their model was able to get a high accuracy for binary as well as multiclass classification tasks, surpassing earlier research on Bengali cyberbullying. **T. H. Teng, et al. [43**] indicated that transformer based transfer learning models outperformed standard machine learning approaches in terms of collecting contextual meaning and boosting prediction capabilities. **S. M. Fati et al. [44]** studied the use of attention-based deep learning architectures paired with Continuous Bag-of-Words (CBOW) embeddings for cyberbullying detection on Twitter. **M. T. Hasan et al. [45]** reviewed cyberbullying detection strategies based on deep learning in detail. To tackle the challenges of standard machine learning methods. Their study pointed out that deep learning methods outperform traditional machine learning techniques owing to better capacity to automatically extract features, grasp context and adapt to different datasets. **B. Ogunleye et al. [46]** studied the use of Large Language Models (LLMs) like BERT and RoBERTa for automatic cyberbullying identification. **P. Yi et al. [47]** presented the LS-CB (Long Session Cyberbullying Detection) framework, a transformer-based technique that aims to identify cyberbullying in long social media interactions, not only in single postings. **A. G. Philipo et al. [48]** developed Hybrid BLOOM-Based Ensemble Model (HBEM) for real-time cyberbullying detection in multilingual, low-resource and code-mixed language settings, focusing on Swahili and Swanglish datasets. HBEM adopts LoRA fine-tuning and ensemble fusion for BLOOM-560M and BLOOM-1B1 models, which can achieve excellent classification performance with computational efficiency. **J. A. Diaz-Garcia et al. [49]** conducted an extensive study on the detection of textual cyber abuse utilizing state-of-the-art natural language processing approaches, including transformer language models and LLMs. **M. Hossain et al. [50]** suggested a hybrid architecture that combines Transformer-XL with bidirectional recurrent neural networks to detect cyberbullying. The model was able to successfully capture long-range contextual relationships and sequential language patterns, leading in better identification of implicit and conversational cyberbullying compared to standalone transformer or recurrent models. **H. Allwaibed et al. [51]** performed a comprehensive review of studies on the identification of cyberbullying in the Arabic language. **D. Kumar et al., [52]** presented a hybrid architecture of DeBERTa and Gated Broad Learning System (GBLS) for English cyberbullying detection.

Table 1: Literature Summary of Text based Social Media Harassment Detection Frameworks.

Detection Category	Attack Type / Target	Dataset / Platform	Model / Framework	Major Findings	Ref.
Text-based	Threats, defamation, identity attacks	Social media platforms	Harassment behavior analysis	Highlighted the growing complexity and prevalence of harmful online interactions	[30]
Text-based	Hate speech and abusive communication	Social media datasets	DL-based contextual toxicity detection	Deep learning improves detection of sarcasm and context-dependent abuse	[31]
Behavioral / Text-based	Cyberbullying perpetration and victimization	Behavioral and cyberbullying studies	Integrative Cyberbullying Theory (ICBT)	Linked cyberbullying with psychological and behavioral outcomes, including depression, anxiety, loneliness, and emotional distress	[32]
Text-based	Implicit aggression and contextual toxicity	Cyberbullying datasets	BullyNet deep learning framework	Adaptive learning systems improve detection of evolving and context-sensitive cyber harassment	[33]
Behavioral	Cyberbullying among adolescents	Interviews with adolescents aged 13–16	Qualitative and empathy-centered cyberbullying analysis	Emphasized empathy, bystander involvement, and teacher–student relationships for cyberbullying prevention	[34]

Text-based	General cyberbullying	Literature and cyberbullying studies	ML-based cyberbullying detection survey	Highlighted the importance of incorporating human behavioral factors into algorithmic detection	[35]
Text-based	General cyberbullying	Multiple cyberbullying datasets	Automated cyberbullying detection pipeline review	Reviewed the complete ML pipeline from data collection to model evaluation	[36]
Text-based / Multilingual	Sarcasm and cyberbullying	Bangla and Chittagonian datasets	Multilingual detection + LIME-based XAI	Improved transparency and detection in multilingual and low-resource environments	[37]
Text-based	Nuanced and context-dependent harassment	Social media datasets	Transformer-based detection strategy	Transformers improved identification of complex and context-dependent harassment	[38]
Text-based	Slang, sarcasm, and evolving cyber harassment	Twitter	DEA-RNN: Elman RNN + Dolphin Echolocation Algorithm	Achieved 90.45% accuracy and outperformed conventional ML and DL approaches	[39]
Text-based / Multilingual	Arabic cyberbullying	Facebook, Twitter, and Instagram	CNN–BiLSTM–GRU + stacked word embeddings	Hybrid deep learning improved feature extraction and Arabic cyberbullying classification	[40]
Text-based / Multilingual	Racism, threats, flaming, and body shaming	Bengali dataset	XLM-RoBERTa	Transformer-based approach achieved the best performance compared with ML and DL methods	[41]
Text-based / Multilingual	Bengali cyberbullying	Bengali dataset	TF-IDF + IHT resampling + ML	Achieved high performance for binary and multiclass cyberbullying classification	[42]
Text-based	Contextual cyberbullying	Contextual cyberbullying datasets	Transfer learning with DistilBERT	Transfer learning improved contextual interpretation and reduced manual feature engineering	[43]
Text-based	Offensive and abusive communication	Twitter	CBOW + attention-based Conv1D-LSTM	Attention mechanisms enhanced contextual cyberbullying detection	[44]
Text-based	General cyberbullying	Multiple datasets / review studies	Deep learning: CNN, RNN, LSTM	Deep learning improved automatic feature extraction and contextual understanding compared with traditional ML	[45]
Text-based	Implicit cyberbullying	Twitter and Formspring datasets	BERT and RoBERTa	RoBERTa demonstrated strong contextual interpretation and cyberbullying detection capability	[46]
Text-based / Conversational	Long-session cyberbullying	Long social media interaction sessions	LS-CB transformer framework	Long-session analysis captured temporal patterns and contextual flow better than single-post approaches	[47]
Text-based / Multilingual	Cyberbullying in low-resource and code-mixed languages	Swahili and Swanglish datasets	Hybrid BLOOM-Based Ensemble Model (HBEM) + LoRA	Improved real-time multilingual and code-mixed cyberbullying detection with	[48]

				computationally efficient fine-tuning	
Text-based	Harassment, grooming, emotional abuse, sextortion, and hate speech	Online and social media text datasets	Transformer language models and LLMs	Contextual language models improved semantic understanding and generalization across diverse forms of cyber abuse	[49]
Text-based	Implicit and conversational cyberbullying	Cyberbullying text datasets	Transformer-XL + Bidirectional RNN	Captured long-range contextual dependencies and sequential patterns, improving implicit and conversational cyberbullying detection	[50]
Text-based / Multilingual	Arabic cyberbullying and abusive language	Arabic social media datasets	Systematic review of multilingual and transformer-based approaches	Identified the need for standardized datasets, multilingual models, and culturally aware annotation	[51]
Text-based / Explainable AI	English cyberbullying and abusive content	HateXplain and other benchmark datasets	DeBERTa + Gated Broad Learning System (GBLS) + XAI	Combined contextual and sentiment-aware learning with token attribution and LIME to improve performance and interpretability	[52]

2.2.2 Image based attack

Image-based attacks represent one of the fastest growing harmful forms of social media harassment, where abusive or manipulated visual content is used to intimidate, humiliate, threaten, or exploit individuals online **(Figure 4)**. The accurate detection and identification of image-based cyber harassment are crucial for protecting users, preventing online abuse, and ensuring a safer digital environment.

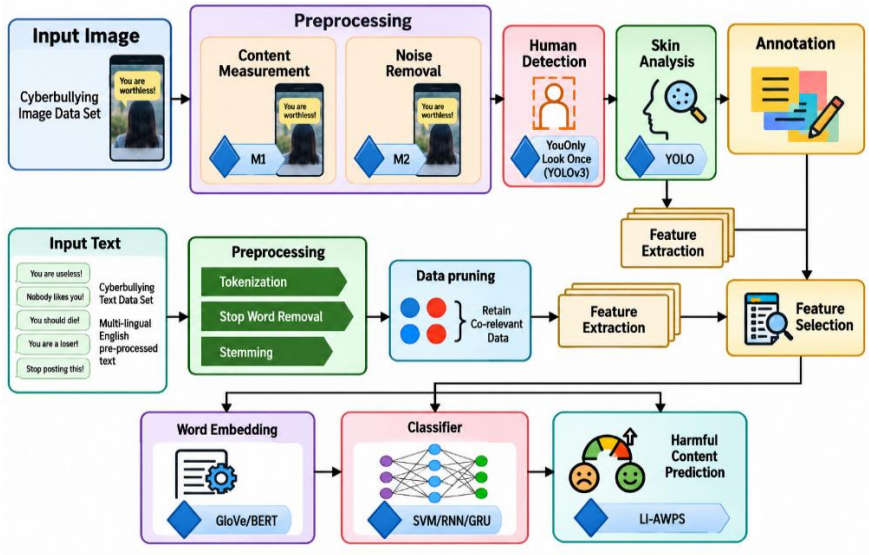

Figure 4. Multimodal framework for cyberbullying and cyberflashing detection. Reproduced with permission from M. Karpagam et al. [53].

S. Pericherla et al., [53] proposed a multimodal cyberbullying detection framework using pretrained deep learning architectures to extract features from textual and image data. **Y. Zhou et al., [54]** investigated multimodal learning for hateful meme detection and emphasized that harmful meanings often emerge from the interaction between images and accompanying text. The authors incorporated an image-captioning process into their detection framework to improve the semantic understanding of visual content. **P. K. Roy et al., [55]** proposed a transfer learning-based deep learning framework for cyberbullying detection. Their findings demonstrated the effectiveness of transfer learning approaches for identifying harmful content, with the best-performing model achieving an accuracy of approximately 89%. **S. Hakimov et al., [56]** developed a multimodal architecture for detecting and classifying misogynistic memes by jointly analyzing textual and visual features. The framework was evaluated in the SemEval-2022 Task 5 Multimedia Automatic Misogyny Identification (MAMI) challenge and achieved the best performance in Task B, which involved identifying categories of misogynistic content, including shaming, stereotyping, objectification, and violence. **M. R. Karim et al., [57]** further investigated multimodal hate speech detection in Bengali social media content by considering both textual information and images embedded in memes. The authors

evaluated several deep learning and transformer-based architectures, including XLM-RoBERTa, ResNet-152, and DenseNet-161. The multimodal fusion of XLM-RoBERTa and DenseNet-161 achieved an F1-score of 0.83, demonstrating that integrating visual and textual information can improve harmful-content detection, particularly in low-resource languages. **S. Schoenebeck et al. [58]** examined the perceived harms associated with different forms of online harassment. Their findings indicated that non-consensual photo sharing, doxxing, and reputational damage were perceived as particularly harmful forms of online harassment. **N. M. Singh et al. [59]** proposed an automated multimodal cyberbullying detection system integrating textual, visual, and audio information to overcome limitations of traditional text-dependent approaches. The framework employe Tuned EfficientNetB0 for image processing, and ACNN for audio feature extraction, followed by a Multi-Modality Decision Fusion Classifier (MMDFC) for final classification. **L. Natrayan et al. [60]** demonstrated that harmful online behaviors frequently occur through memes, images, and multimedia content, highlighting the inadequacy of relying solely on textual analysis for moderation systems. **R. Suhas Bharadwaj et al. [61]** proposed a multimodal hybrid framework for cyberbullying detection. The study employed Convolutional Neural Networks (CNNs) for emoji interpretation and an Extreme Learning Machine (ELM) classifier for final cyberbullying prediction following preprocessing and feature extraction.

Table 2: Literature Summary of Image based Social Media Harassment Detection Frameworks

Detection Category	Attack Type / Target	Dataset / Platform	Major Findings	Ref.
Image + Text-based	Image- and text-based cyberbullying	Social media multimodal data	Integration of textual and visual features improved cyberbullying identification compared with unimodal approaches	[54]
Image + Text-based	Hateful and abusive memes	Hateful Memes Challenge dataset	Improved semantic understanding of image–text relationships and demonstrated the importance of cross-modal reasoning	[55]
Text-based / Cyberbullying	General cyberbullying and harmful content	Cyberbullying datasets	Transfer learning approaches improved cyberbullying detection, with the best-performing model achieving approximately 89% accuracy	[56]
Image + Text-based	Misogynistic memes, shaming, stereotyping, objectification, and violence	SemEval-2022 Task 5 MAMI dataset	Joint visual and textual representations effectively detected and categorized misogynistic meme content	[57]
Image + Text-based / Multilingual	Hate speech in Bengali memes	Bengali social media and meme datasets	XLM-RoBERTa and DenseNet-161 multimodal fusion achieved an F1-score of 0.83	[58]
Image-based / Harassment Analysis	Non-consensual photo sharing, doxxing, and reputational attacks	Online and social media platforms	Non-consensual photo sharing, doxxing, and reputational damage were perceived as particularly harmful forms of online harassment	[59]
Image + Text + Audio	Multimedia cyberbullying	Multimodal datasets	Integration of textual, visual, and audio features substantially improved detection, achieving 98.23% accuracy	[60]
Image / Multimedia-based	Harmful content in memes, images, and multimedia	Social media multimedia content	Highlighted the inadequacy of text-only moderation and emphasized the importance of analyzing multimedia content	[61]
Multimodal	Cyberbullying involving emojis and other multimodal cues	Social media datasets	CNN-based emoji interpretation combined with ELM classification supported multimodal cyberbullying prediction	[62]

2.2.3 Video based attack detection

Videos relating to harassment include cyberbullying videos, abuse videos broadcast live, deepfake movies, revenge porn, films depicting violence, hate propaganda, sexually explicit material, altered videos defaming someone, and videos that misrepresent the context. Thus, an accurate video-based cyber harassment detection is an essential research area to effectively detect the harmful contents **(Figure 5)**.

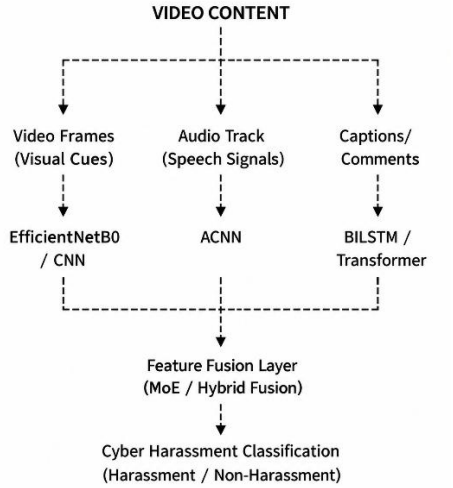

Figure 5. Multimodal framework for video-based cyberbullying detection integrating textual, visual, and audio modalities through a Multi-Modality Decision Fusion Classifier (MMDFC). Reproduced with permission from N. M. Singh et al. [59].

N. M. Singh et al. [59] proposed an automated multimodal cyberbullying detection system that combines text, visual and audio information. **C. W. Jo et al. [63]** used multimedia analysis to the more general identification of hazardous material on video-sharing sites. The research studied 19422 YouTube videos with numerous video frames, thumbnails and textual information, and explored the possibility of multimodal Large Language Models (MLLMs) as alternative annotators. The results highlighted the promise of MLLMs for detection and classification of different types of harmful video material, while also showing the challenges of defining, annotating and discriminating across various categories of online damage. **R. S. Bharadwaj et al. [62]** proposed a hybrid multimodal framework for cyberbullying detection by exploiting multiple social media cues. **C. Catalano et al. [64]** explored the capabilities of multimodal Large Language Models to detect cyberbullying directly from images and videos. **A. Yadav et al. [65]** presented an interpretable agentic paradigm for multimodal hate-video analysis with explicit evidence attribution. Instead of treating multimodal information as a single fused representation, the proposed framework analyzed visual, audio, textual and contextual signals as independent sources of evidence.

Table 3: Literature Summary of video based Social Media Harassment Detection Frameworks

Detection Category	Attack Type / Target	Dataset / Platform	Model / Framework	Major Findings	Ref.
Video + Image + Audio + Text	Multimodal cyberbullying	Multimodal social media datasets	Automated multimodal cyberbullying detection framework	Integration of textual, visual, and audio information improves cyberbullying detection beyond traditional text-dependent approaches	[60]
Multimedia	Harassment through images, memes, videos, and other multimedia content	Social media multimedia content	Multimedia-based cyber harassment analysis	Highlighted the limitations of text-only moderation and emphasized the need to analyze multimedia content for comprehensive harassment detection	[61]
Multimodal	Cyberbullying involving multiple social media cues	Social media datasets	Hybrid multimodal cyberbullying detection framework	Incorporating multiple social media cues improves the identification of cyberbullying compared with unimodal approaches	[62]
Video + Image + Text	Information-related harm, hate and harassment, addictive content, clickbait, sexual harm, and physical harm	19,422 YouTube videos	Harm taxonomy + Multimodal Large Language Models (MLLMs)	Multimodal analysis of video frames, thumbnails, and textual metadata demonstrated the potential of MLLMs for identifying and categorizing diverse harmful video content	[63]
Video + Image-based Multimodal	Revenge pornography, happy slapping, racism, and body shaming	5,000+ multimedia samples from Telegram, Reddit, and X	Zero-shot multimodal LLMs: LLaMA 4, Gemma 3, and Gemini	Multimodal LLMs demonstrated competitive performance in detecting explicit and subtle online aggression; Gemma 3-12B showed the most stable overall performance	[64]

Video-based Multimodal	Explicit and implicit hate in online videos	HateMM, ImpliHateVid, and YTHate (200 YouTube videos)	Interpretable agentic multimodal framework with explicit evidence attribution	Independent analysis of visual, audio, textual, and contextual evidence improved transparency, interpretability, and auditability of harmful-video detection	[65]

4. CHALLENGES AND RESEARCH NEEDS

Despite great progress, there are still various problems in cyber harassment detection. Sarcasm and implicit toxicity, comedy and coded language detection remains a challenge, especially in bilingual and cross-cultural environments. Most of the available models are trained on English datasets and they are not effective for low resource languages. Model reliability and fairness are affected by dataset imbalance, inconsistent annotations, and poor context information, Improved multimodal and cross-modal analysis to address new threats such as deepfakes and AI-generated multimedia. Also, computational complexity, limited explainability, adversarial assaults, privacy issues and real-time deployment limits are still limiting the development of strong and scalable cyber harassment detection systems.

5. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Social media and online communication are rapidly growing and cyber-harassment is becoming an increasingly relevant topic. Today's artificial intelligence and machine learning methods hold great promise for detecting and decreasing dangerous information online. However, the efficiency of existing systems remains limited due to issues such as contextual ambiguity, multilingual communication, developing forms of harassment, privacy concerns, model bias, and identification of coordinated assaults. Hence, future cyber-harassment detection systems should be accurate, adaptable, transparent and user-aware.

In the future, more studies on Explainable AI (XAI) methods such as LIME and SHAP may be conducted to increase the transparency and interpretability of automated moderation judgments. Federated learning may be used to train models in a privacy-preserving manner, allowing collaborative learning without directly exchanging sensitive user data. Graph neural networks (GNNs)may be used to identify concerted harassment by examining relationships and interactions among users. Furthermore, large language models (LLMs) and lightweight transformer-based models are capable of enhancing contextual, multilingual and conversational toxicity detection.

Future study should also focus on the identification of deepfakes and material created by AI using multimodal learning, emotion-aware cyberbullying detection and cross-platform harassment analysis. It is important to create fair, bias-resistant, and privacy-preserving AI frameworks that can function reliably across languages, cultures, and user groups. In sum, future research should concentrate on the trade-off between effective harassment detection and privacy, fairness, transparency, and freedom of speech to enable safer and more responsible online environments.

6. REFERENCES

[1] Kaplan, A.M. and Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), pp.59-68.

[2] Kietzmann, J.H., Hermkens, K., McCarthy, I.P. and Silvestre, B.S., 2011. Social media? Get serious! Understanding the functional building blocks of social media. *Business horizons*, *54*(3), pp.241-251.

[3] Ferrara, E., Varol, O., Davis, C., Menczer, F. and Flammini, A., 2016. The rise of social bots. *Communications of the ACM*, *59*(7), pp.96-104.

[4] Vosoughi, S., Roy, D. and Aral, S., 2018. The spread of true and false news online. *science*, *359*(6380), pp.1146-1151.

[5] Kowalski, R.M., Giumetti, G.W., Schroeder, A.N. and Lattanner, M.R., 2014. Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin*, *140*(4), p.1073.

[6] Patchin, J.W. and Hinduja, S., 2015. Measuring cyberbullying: Implications for research. *Aggression and violent behavior*, *23*, pp.69-74.

[7] Vidgen, B. and Derczynski, L., 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, *15*(12), p.e0243300.

[8] Schmidt, A. and Wiegand, M., 2017, April. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1-10).

[9] Wang, H., Kumar, R., Pattanaik, A., Kumar, R., Khawaf Aljaberi, A.S.O. and Abass, M.A., 2025. Computational methods and artificial intelligence-based modeling of magnesium alloys: a systematic review of machine learning, deep learning, and data-driven design and optimization approaches. *Frontiers in Materials*, *12*, p.1645227.

[10] Vidgen, B. and Derczynski, L., 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, *15*(12), p.e0243300.

[11] Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., 2016. *Deep learning* (Vol. 1, No. 2, pp. 1-800). Cambridge: MIT press.

[12] Kim, Y., 2014, October. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1746-1751).

[13] Lipton, Z.C., Kale, D.C., Elkan, C. and Wetzel, R., 2015. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677*.

[14] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, *9*(8), pp.1735-1780.

[15] Chen, X., He, J., Wu, X., Yan, W. and Wei, W., 2020. Sleep staging by bidirectional long short-term memory convolution neural network. *Future Generation Computer Systems*, *109*, pp.188-196.

[16] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language*

technologies, volume 1 (long and short papers) (pp. 4171-4186).

[17] Ruder, S., Peters, M.E., Swayamdipta, S. and Wolf, T., 2019, June. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials* (pp. 15-18).

[18] Nova, F.F., DeVito, M.A., Saha, P., Rashid, K.S., Roy Turzo, S., Afrin, S. and Guha, S., 2020, October. Understanding how marginalized hijra in bangladesh navigate complex social media ecosystem. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 353-358).

[19] Vogels, E.A., Gelles-Watnick, R. and Massarat, N., 2022. Teens, social media and technology 2022. *Pew Research Center*, *10*, pp.1-30.

[20] Schoenebeck, S., Batool, A., Do, G., Darling, S., Grill, G., Wilkinson, D., Khan, M., Toyama, K. and Ashwell, L., 2023, April. Online harassment in majority contexts: Examining harms and remedies across countries. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1-16).

[21] Ray, G., McDermott, C.D. and Nicho, M., 2024. Cyberbullying on social media: Definitions, prevalence, and impact challenges. *Journal of cybersecurity*, *10*(1), p.tyae026.

[22] Saha, R., Ahlawat, S., Akram, U., Jangbahadur, U., Dhaigude, A.S., Sharma, P. and Kumar, S., 2024. Online abuse: a systematic literature review and future research agenda. *International Journal of Conflict Management*, *35*(5), pp.887-917.

[23] Joseph, J., 2026. Digital Attacks and Online Harassment as a New Form of Violence Against Women: Ethical and Legal Considerations. In *Policy, Prevention, and Structural Responses to Violence Against Women* (pp. 161-188). IGI Global Scientific Publishing.

[24] Milyane, T.M., Rohimakumullah, M.A.A. and Fariza, M.R., 2025. Exploring Adolescents' Understanding of Cyberbullying on Social Media Through Information Mining. *Howard Journal of Communications*, *36*(4), pp.473-487.

[25] Bachmann, I., 2024. Resilience, support, and feminist counterpublics in online debates of gender-based violence in Latin America. In *New Digital Feminist Interventions* (pp. 32-47). Routledge.

[26] Pavón Pérez, Á., Farrell, T., De Kock, C., Jurasz, O., Nozza, D. and Fernandez, M., 2026. Still unsafe: what's holding us back on online safety for women. *AI and Ethics*, *6*(2), p.246.

[27] Yi, P. and Zubiaga, A., 2023. Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*, *36*, p.100250.

[28] Akwa, G., 2026. Technological transitions in artificial intelligence-driven cyberbullying mitigation on social media: a systematic review. *AI and Ethics*, *6*(2), p.221.

[29] Diaz-Garcia, J.A. and Carvalho, J.P., 2025. A Literature Review of Textual Cyber Abuse Detection Using Cutting-Edge Natural Language Processing Techniques: Language Models and Large Language Models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *15*(3), p.e70029.

[30] Johri, A. and Singh, R., 2024. Cyberbullying on Social Networks–A Crime in India. *International Journal of Indian Psychology*, *12*(3).

[31] Alabdulwahab, A., Haq, M.A. and Alshehri, M., 2023. Cyberbullying detection using machine learning and deep learning. *International Journal of Advanced Computer Science and Applications*, *14*(10).

[32] Barlett, C.P., Gentile, D.A., Khoo, A., Kowalski, R. and Waasdorp, T.E., 2026. The Predictors and Consequences of Cyberbullying Perpetration and Cyber-Victimization: The Development of a New Cyberbullying Theory. *Journal of Adolescence*, *98*(1), pp.69-81.

[33] Sree, S.S. and Joseph, L.N., 2025. Revolutionizing cyber-bullying detection with the bullynet deep learning framework. *Int Res J Multidiscip Technovation*, *7*(2), pp.38-49.

[34] Pyżalski, J., Plichta, P., Szuster, A. and Barlińska, J., 2022. Cyberbullying characteristics and prevention—what can we learn from narratives provided by adolescents and their teachers?. *International journal of environmental research and public health*, *19*(18), p.11589.

[35] Kim, S., Razi, A., Stringhini, G., Wisniewski, P.J. and De Choudhury, M., 2021. A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), pp.1-34.

[36] Elsafoury, F., Katsigiannis, S., Pervez, Z. and Ramzan, N., 2021. When the timeline meets the pipeline: A survey on automated cyberbullying detection. *IEEE access*, *9*, pp.103541-103563.

[37] Mahmud, T., Ptaszynski, M. and Masui, F., 2024. Exhaustive study into machine learning and deep learning methods for multilingual cyberbullying detection in bangla and chittagonian texts. *Electronics*, *13*(9), p.1677.

[38] Abdullah, A., Ullah, F., Hafeez, N., Latif, I., Sidorov, G., Riveron, E.F. and Gelbukh, A., 2025. Cyberbullying detection on social media using machine learning techniques. *Computación y Sistemas*, *29*(3).

[39] Murshed, B.A.H., Abawajy, J., Mallappa, S., Saif, M.A.N. and Al-Ariki, H.D.E., 2022. DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. *IEEE Access*, *10*, pp.25857-25871.

[40] Daraghmi, E.Y., Qadan, S., Daraghmi, Y.A., Yousuf, R., Cheikhrouhou, O. and Baz, M., 2024. From text to insight: An integrated cnn-bilstm-gru model for arabic cyberbullying detection. *IEEE Access*, *12*, pp.103504-103519.

[41] Sihab-Us-Sakib, S., Rahman, M.R., Forhad, M.S.A. and Aziz, M.A., 2024. Cyberbullying detection of resource constrained language from social media using transformer-based approach. *Natural Language Processing Journal*, *9*, p.100104.

[42] Akhter, A., Acharjee, U.K., Talukder, M.A., Islam, M.M. and Uddin, M.A., 2023. A robust hybrid machine learning model for Bengali cyber bullying detection in social media. *Natural Language Processing Journal*, *4*, p.100027.

[43] Teng, T.H. and Varathan, K.D., 2023. Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access*, *11*, pp.55533-55560.

[44] Fati, S.M., Muneer, A., Alwadain, A. and Balogun, A.O., 2023. Cyberbullying detection on twitter using deep learning-based attention mechanisms and continuous Bag of words feature extraction. *Mathematics*, *11*(16), p.3567.

[45] Hasan, M.T., Hossain, M.A.E., Mukta, M.S.H., Akter, A., Ahmed, M. and Islam, S., 2023. A review on deep-learning-based cyberbullying detection. *Future Internet*, *15*(5), p.179.

[46] Ogunleye, B. and Dharmaraj, B., 2023. The use of a large language model for cyberbullying detection. *Analytics*, *2*(3), pp.694-707.

[47] Yi, P. and Zubiaga, A., 2023, April. Learning like human annotators: Cyberbullying detection in lengthy social media sessions. In *Proceedings of the ACM web conference 2023* (pp. 4095-4103).

[48] Philipo, A.G., Sarwatt, D.S., Ding, J. and Ning, H., 2026. A Lightweight Ensemble Model for Real-Time Cyberbullying Detection in Low-Resource and Code-Mixed Contexts. *Authorea Preprints*.

[49] Diaz-Garcia, J. Angel, Carlos Fernandez-Basso, Jesica Gómez-Sánchez, Karel Gutiérrez-Batista, M. Dolores Ruiz, and Maria J. Martin-Bautista. "A fuzzy-based approach for cyberbullying analysis." In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 317-328. Cham: Springer International Publishing, 2022.

[50] Hossain, M.M., Hossain, M.S., Hossain, M.S., Mridha, M.F., Safran, M., Alfarhood, S. and Che, D., 2025. Fusing Transformer-XL with bi-directional recurrent networks for cyberbullying detection. *PeerJ Computer Science*, *11*, p.e2940.

[51] Allwaibed, H., Anbar, M., Manickam, S. and Bintang, A., 2025. Cyberbullying detection approaches for Arabic texts: a systematic literature review. *Frontiers in Artificial Intelligence*, *8*, p.1666349.

[52] Kumar, D., 2025. A Hybrid DeBERTa and Gated Broad Learning System for Cyberbullying Detection in English Text. *arXiv preprint arXiv:2506.16052*.

[53] Karpagam, M., Naveenkumar, N., Panguluri, V., Hanuman, C.R.S., Usharani, R., Priya, S. and Sharma, P.C., 2025. An effective cyberbullying-flashing identification on whatsapp using PTS-GReLU-GRU with harmful level prediction. *Scientific Reports*.

[54] Pericherla, S. and Egambaram, I., 2021. Cyberbullying detection on multi-modal data using pre-trained deep learning architectures. *Ingeniería Solidaria*, *17*(3), pp.1-20.

[55] Zhou, Y., Chen, Z. and Yang, H., 2021, July. Multimodal learning for hateful memes detection. In *2021 IEEE International conference on multimedia & expo workshops (ICMEW)* (pp. 1-6). IEEE.

[56] Roy, P.K. and Mali, F.U., 2022. Cyberbullying detection using deep transfer learning. *Complex & Intelligent Systems*, *8*(6), pp.5449-5467.

[57] Hakimov, S., Cheema, G.S. and Ewerth, R., 2022, July. TIB-VA at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 756-760).

[58] Karim, M.R., Dey, S.K., Islam, T., Shajalal, M. and Chakravarthi, B.R., 2022, November. Multimodal hate speech detection from bengali memes and texts. In *International Conference on Speech and Language Technologies for Low-resource Languages* (pp. 293-308). Cham: Springer International Publishing.

[59] Schoenebeck, S., Batool, A., Do, G., Darling, S., Grill, G., Wilkinson, D., Khan, M., Toyama, K. and Ashwell, L., 2023, April. Online harassment in majority contexts: Examining harms and remedies across countries. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1-16).

[60] Singh, N.M. and Sharma, S.K., 2024. An efficient automated multi-modal cyberbullying detection using decision fusion classifier on social media platforms. *Multimedia Tools and Applications*, *83*(7), pp.20507-20535.

[61] Natrayan, L., Sirisha, J. and JV, R.K., 2025, September. A Multi-Model Ensemble Learning Approach for Cyberbullying Detection in Social Networks. In *2025 IEEE International Conference on Advances in Computing Research On Science Engineering and Technology (ACROSET)* (pp. 1-6). IEEE.

[62] Suhas Bharadwaj, R., Kuzhalvaimozhi, S. and Vedavathi, N., 2022. A novel multimodal hybrid classifier based cyberbullying detection for social media platform. In *Proceedings of the Computational Methods in Systems and Software* (pp. 689-699). Cham: Springer International Publishing.

[63] Jo, C.W. and Wojcieszak, M., 2024. Harmful youtube video detection: A taxonomy of online harm and mllms as alternative annotators. *arXiv preprint arXiv:2411.05854*.

[64] Catalano, C., Chezzi, A., Lupo, S., Catalano, A.A., Vadacca, R. and Mainetti, L., 2026. Multimodal Cyberbullying Detection on Social Media Using LLMs: A Comparative Study. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *22*(1), pp.1-21.

[66] Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso, P., Lees, A. and Sorensen, J., 2022, July. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 533-549).

[67] Chen, L. and Chou, H.W., 2022, July. RIT boston at semeval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from CLIP model and data-centric AI principle. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 636-641).

[68] Paraschiv, A., Dascalu, M. and Cercel, D.C., 2022, July. UPB at semeval-2022 task 5: Enhancing UNITER with image sentiment and graph convolutional networks for multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 618-625).

[69] Yoder, M.M., Ng, L.H.X., Brown, D. and Carley, K.M., 2022, December. How hate speech varies by target identity: A computational analysis. In *Proceedings of the 26th*

Conference on Computational Natural Language Learning (CoNLL) (pp. 27-39).

[70] Fetahi, E., Susuri, A., Hamiti, M., Kastrati, Z., Canhasi, E. and Misini, A., 2025. Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI. *Social Network Analysis and Mining*, *15*(1), p.82.

[71] Chhabra, A. and Vishwakarma, D., 2024. Mhs-stma: Multimodal hate speech detection via scalable transformer-based multilevel attention framework. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[72] Xing, R., Chai, Q., Ma, J., Tao, J., Wang, P., Zhang, S., Wang, X. and Wang, H., 2026. Is AI Ready for Multimodal Hate Speech Detection? A Comprehensive Dataset and Benchmark Evaluation. *arXiv preprint arXiv:2603.21686*.

[73] Islam, M.R., Rahman, M.M., Hossain, M.S. and Hasan, M.K., 2024. Cyberbullying and online harassment detection on social media using machine learning and sentiment analysis techniques. Journal of Information and Communication Technology, 23(2), pp.145–162.