Chapter 1

# Artificial Intelligence-Based Cyberbullying Detection and Prevention: Deep Learning Architectures, Multimodal Analysis, Ethical Challenges, and Future Directions

**Sara Naghib Zadeh**[1]

**Zühre Aydin**[2]

**Abstract**

The rapid growth of social media and digital communication platforms has significantly increased the prevalence of cyberbullying, online harassment, and hate speech. Due to the large volume and dynamic nature of online content, manual monitoring has become insufficient, leading to the growing use of artificial intelligence (AI)-based detection and prevention systems. Cyberbullying is not only a technical problem but also a major social and psychological challenge with serious consequences for individuals and online communities.

This paper presents a comprehensive review of AI-based cyberbullying detection approaches, focusing on machine learning, deep learning, and multimodal analysis techniques. The study examines traditional machine learning methods alongside advanced deep learning architectures such as CNN, RNN, LSTM, hybrid CNN–LSTM models, and transformer-based models including BERT. In addition, the paper discusses multimodal systems that combine textual, visual, and sentiment-based analysis to improve the detection of implicit and complex harmful content.

The study also addresses important challenges such as adversarial attacks, linguistic manipulation, dataset imbalance, algorithmic bias, privacy concerns, and ethical issues related to automated moderation systems. Furthermore,

1 Dr. Lecture, Halic University, Vocational School, Department of Computer Programming, ORCID: 0009-0005- 6959-1165.

2 Halic University, Vocational School, Department of Big Data Analytics, ORCID: 0009-0009-4523-9406

https://doi.org/10.58830/ozgur.pub1351.c5534

future directions involving explainable AI, predictive moderation systems, and human–AI collaborative frameworks are explored.

The findings indicate that although AI-based systems have significantly improved cyberbullying detection performance, achieving a balance between technical accuracy, fairness, transparency, and freedom of expression remains a major challenge. Future progress in this field will require interdisciplinary approaches that integrate advanced AI technologies with ethical and humancentered moderation strategies.

## 1. Cyberbullying Development in the Digital Environment and Detection Approaches

With the rapid expansion of communication technologies and the pervasive influence of social media in everyday life, human interaction has increasingly shifted from physical environments to digital ecosystems. This transformation has fundamentally changed the nature of social communication, enabling instant information exchange, global connectivity, and unprecedented access to digital platforms. However, alongside these advantages, the same environment has also facilitated the emergence and rapid spread of harmful online behaviors such as cyberbullying, online harassment, and hate speech.

One of the most critical factors contributing to cyberbullying is anonymity in online environments. Anonymity reduces accountability and psychological inhibition, allowing individuals to express aggressive behaviors that they would typically avoid in face-to-face interactions. Research has consistently shown that anonymity plays a central role in increasing the likelihood and severity of online aggression . In addition, the perceived distance between users in digital environments further amplifies disinhibition effects, making cyberbullying more frequent and less controllable compared to traditional bullying (Al-Ajlan & Ykhlef, 2018; Al-Dabet et al., 2023).

From a psychological and social perspective, cyberbullying has significant and long-lasting consequences, especially among adolescents. This group is particularly vulnerable due to their developmental stage, higher dependency on peer validation, and intensive use of social media platforms. Exposure to repeated online harassment can lead to serious mental health issues such as anxiety, depression, reduced self-esteem, social withdrawal, and in extreme cases, suicidal ideation .Unlike traditional bullying, the digital nature of cyberbullying ensures continuous exposure, as harmful content can persist online indefinitely and be accessed repeatedly (Aldreabi, 2024; Alabdulwahab et al., 2023).

Furthermore, cyberbullying is not an isolated individual behavior but a complex socio-technical phenomenon influenced by platform design, user

interaction patterns, and algorithmic content distribution. Recommendation systems, content virality mechanisms, and social reinforcement loops can unintentionally amplify harmful content, thereby increasing its visibility and impact. As illustrated in Figure 1, the rapid expansion of digital communication environments, combined with anonymity and large-scale social interaction, has accelerated the spread of cyberbullying and increased the need for AI-based automated detection systems.

In response to these challenges, automated detection systems based on machine learning have been introduced as a fundamental solution for identifying harmful content in large-scale social media environments. These systems are capable of processing massive volumes of textual data and detecting linguistic patterns associated with offensive, abusive, or threatening content. Studies demonstrate that machine learning algorithms perform effectively in classifying cyberbullyingrelated content, particularly in platforms such as Twitter where data volume and velocity are high (Muneer & Fati, 2020).

The integration of Natural Language Processing (NLP) techniques with machine learning further enhances system performance by enabling more accurate feature extraction from textual data. NLP techniques such as tokenization, sentiment analysis, and word embedding representations play a crucial role in improving classification accuracy and contextual understanding (LeCun et al., 2015).

Despite their effectiveness, traditional machine learning approaches such as Support Vector Machines (SVM), Naïve Bayes, and Random Forest have inherent limitations. These models rely heavily on manual feature engineering and often fail to capture contextual dependencies, sarcasm, and semantic complexity in language. As a result, their performance decreases significantly in realworld social media environments where language is dynamic, informal, and context-dependent (Aqeel & Kamble, 2022).

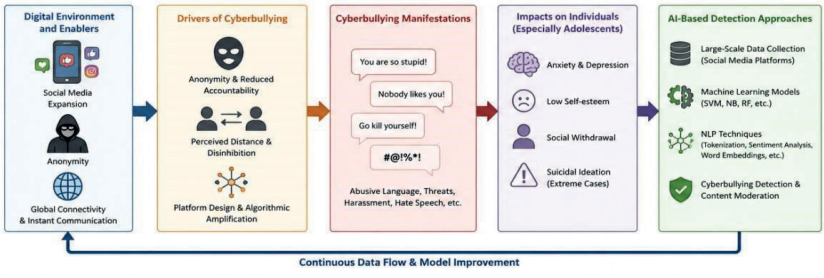


***Figure 1. Evolution of Cyberbullying in Digital Environments and AI-Based Detection Framework***

## 2. Deep Learning Architectures for Cyberbullying Detection

Deep learning represents a significant advancement in machine learning, enabling the modeling of complex nonlinear relationships through multi-layer neural network architectures. These models are capable of automatically extracting hierarchical representations from raw data, eliminating the need for manual feature engineering and significantly improving classification performance in complex tasks (LeCun et al., 2015).

In the field of cyberbullying detection, deep learning has become the dominant methodological approach due to its ability to handle large-scale, unstructured, and noisy textual data. Table 1 provides a comparative overview of the most commonly used deep learning architectures in cyberbullying detection, highlighting their strengths, limitations, and application domains.

### 2.1. CNN-Based Models

Convolutional Neural Networks (CNNs) are widely used for text classification tasks due to their ability to capture local patterns and spatial relationships between words. In cyberbullying detection, CNNs are particularly effective in analyzing short and informal texts commonly found on social media platforms. These models can identify local semantic patterns such as offensive phrases, repeated word structures, and contextual word groupings. Wang et al. demonstrate that CNN-based models achieve high accuracy in detecting abusive and offensive language in short social media texts (Wang et al., 2014).

### 2.2. RNN and LSTM-Based Models

Recurrent Neural Networks (RNNs) are designed to capture sequential dependencies in text data, making them suitable for modeling sentence-level context. However, traditional RNNs suffer from vanishing gradient problems, which limit their ability to learn long-term dependencies. To address this limitation, Long Short-Term Memory (LSTM) networks were introduced, enabling more effective learning of long-range contextual relationships in text sequences (Meta Transparency Center, 2024; Aldreabi & Blackburn, 2023).

LSTM-based models are particularly effective in detecting subtle linguistic cues such as implicit aggression and contextual negativity, which are often overlooked by simpler models.

### 2.3. Hybrid Architectures

To leverage the strengths of both CNN and RNN architectures, hybrid models such as CNN–LSTM and CNN–LRCN have been proposed. These

models combine local feature extraction (CNN) with sequential learning (RNN/LSTM), resulting in improved classification performance. Empirical studies show that hybrid architectures significantly outperform traditional machine learning methods in cyberbullying and toxic content detection tasks (Meta Transparency Center, 2024; SJ & Cho, 2020).

Hybrid models are particularly effective in handling social media data, which is typically short, noisy, and linguistically inconsistent.

### 2.4. Transformer-Based Models

The introduction of transformer architectures has revolutionized natural language processing. Models such as BERT (Bidirectional Encoder Representations from Transformers) provide deep contextual understanding by analyzing text in both forward and backward directions simultaneously. This bidirectional learning mechanism enables more accurate semantic representation of language (European Union, 2018).

Pre-trained BERT models have shown strong performance in detecting cyberbullying, hate speech, and offensive content. These models are especially effective in handling complex linguistic phenomena such as sarcasm, irony, and implicit aggression, which are challenging for traditional architectures (DataTurks, 2018; Mozafari et al., 2019).

***Table 1. Comparison of Deep Learning Architectures for Cyberbullying Detection***

| Architecture Type | Key Idea | Strengths | Limitations | Application in Cyberbullying Detection | Representative References |
|---|---|---|---|---|---|
| **CNN-Based Models** | Extract local features using convolution filters over text | Efficient feature extraction, strong for short texts, captures local patterns (e.g., offensive phrases) | Limited ability to model long-range dependencies | Detects abusive words, offensive phrases, and short toxic posts in social media | [109] |
| **RNN / LSTM-Based Models** | Models sequential dependencies in text data | Captures context and temporal dependencies, effective for sentence-level understanding | Vanishing gradient (RNN), higher computational cost (LSTM) | Detects implicit aggression, contextual negativity, and sequential linguistic cues | [31], [52] |
| **Hybrid Models (CNN–LSTM, CNN–LRCN)** | Combines CNN for feature extraction and RNN/LSTM for sequence modeling | Higher accuracy, captures both local and global context | More complex architecture, higher training cost | Effective for noisy social media text and mixed linguistic patterns | [31], [97] |

| | | | | | |
|---|---|---|---|---|---|
| **Transformer-Based Models (e.g., BERT)** | Uses self-attention mechanism for bidirectional context learning | Captures deep contextual meaning, handles sarcasm and implicit hate speech well | Computationally expensive, requires large-scale pretraining | State-of-the-art performance in cyberbullying, hate speech, and offensive language detection | [17], [73], [58] |

## 3. AI-Based Security Systems and Future Perspectives in Cyberbullying Prevention

The rapid expansion of digital communication platforms has created new challenges for ensuring user safety in online environments. Social media platforms, messaging applications, and online communities generate vast amounts of user-generated content every day, making manual monitoring increasingly difficult. As a result, artificial intelligence (AI)-based security systems have emerged as essential tools for supporting content moderation, risk assessment, and cyberbullying prevention efforts (Hussain et al., 2018; Wang et al., 2021).

AI-based security systems are designed to identify, analyze, and respond to harmful online behaviors in real time. Unlike traditional moderation approaches that rely heavily on human reviewers, these systems can continuously monitor large-scale digital environments and assist platforms in detecting potentially harmful interactions more efficiently. Automated moderation tools help reduce response times and improve the scalability of online safety operations (Aqeel & Kamble, 2022).

One of the most important developments in this area is the integration of real-time content moderation mechanisms. Modern security frameworks can automatically flag suspicious content, prioritize high-risk interactions for human review, and support platform administrators in enforcing community guidelines. Such systems contribute to creating safer online environments while reducing the workload of moderation teams (Hussain et al., 2018).

Beyond content moderation, AI technologies are increasingly being used for behavioral analysis. Rather than focusing solely on individual messages, advanced security systems can evaluate patterns of user behavior, interaction frequency, communication networks, and historical activity records. This broader perspective enables the identification of recurring harmful behaviors and potential risk factors associated with cyberbullying incidents (Wang et al., 2021).

Future security frameworks are expected to become more predictive rather than purely reactive. Predictive intelligence systems aim to identify behavioral indicators that may signal the emergence of harmful online interactions before significant damage occurs. By combining behavioral analytics, social network analysis, and large-scale data processing, these systems may provide early warning mechanisms for cyberbullying prevention (Jahan & Oussalah, 2023; Aqeel & Kamble, 2022).

Another important trend is the development of multimodal AI systems. Online communication is no longer limited to text-based interactions. Images, videos, emojis, GIFs, voice recordings, and multimedia content increasingly shape user communication patterns. Consequently, future cyberbullying prevention frameworks are expected to incorporate multiple data sources to achieve a more comprehensive understanding of online behavior. Studies indicate that multimodal analysis can improve the identification of harmful content that may not be explicitly expressed through text alone (Wang et al., 2022).

Large-scale social media platforms also require adaptive security infrastructures capable of responding to rapidly changing communication patterns. New forms of online harassment continuously emerge through evolving slang, coded language, and platform-specific behaviors. Therefore, future AI security systems must be capable of continuous learning and adaptation to maintain effectiveness in dynamic digital environments (Weimann & Masri, 2020).

Despite these advancements, the deployment of AI-based security systems raises significant ethical and regulatory concerns. Automated moderation systems may produce biased outcomes, disproportionately affect certain user groups, or incorrectly restrict legitimate forms of expression. Questions regarding transparency, accountability, privacy protection, and freedom of speech remain central challenges in the development of trustworthy AI systems (Floridi & Cowls, 2019; Hosseini et al., 2017).

To address these concerns, researchers increasingly advocate the adoption of explainable AI principles. Explainable systems provide greater transparency regarding how moderation decisions are made, allowing users, platform administrators, and policymakers to better understand automated outcomes. Such transparency is essential for building trust and ensuring responsible AI governance (Hosseini et al., 2017; Selbst et al., 2019).

Several Explainable AI (XAI) techniques have been proposed to improve transparency in cyberbullying detection systems. Methods such as SHAP

(SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) help identify the most influential features behind model predictions. In transformer-based architectures, attention-based explanation methods further enhance interpretability by highlighting words and contextual patterns that contribute to classification outcomes. These techniques improve transparency, fairness assessment, error analysis, and user trust in AI-driven moderation systems (Arrieta et al., 2020; Aldreabi & Blackburn, 2023).

Looking ahead, AI-based security systems are expected to evolve into integrated digital safety ecosystems that combine real-time monitoring, predictive risk assessment, multimodal analysis, and human oversight. These systems will play a critical role in supporting cyberbullying prevention strategies while balancing technological effectiveness with ethical responsibility and user rights (Diaz-Garcia & Carvalho, 2024; Wang et al., 2021).

Large Language Models (LLMs) and foundation models have recently emerged as powerful tools for cyberbullying detection and content moderation. Models such as GPT, Llama, PaLM, and Claude are capable of understanding complex linguistic structures, contextual meaning, implicit aggression, sarcasm, and culturally dependent expressions. Unlike traditional machine learning systems, LLMs can perform zero-shot and few-shot classification, reducing the need for taskspecific training datasets. Furthermore, foundation models can support multilingual moderation and adaptive content analysis across different online platforms. However, concerns regarding hallucinations, bias, explainability, computational cost, and ethical governance remain significant challenges for their large-scale deployment in cyberbullying prevention systems (Brown et al., 2020; Diaz-Garcia & Carvalho, 2024; Ziems et al., 2023).

***Table 2. Emerging AI Approaches in Cyberbullying Prevention***

| Approach | Main Benefit | Key Challenge |
|---|---|---|
| Predictive AI | Early risk detection | False predictions |
| Multimodal AI | Detects text and visual abuse | High complexity |
| Explainable AI (XAI) | Improves transparency and trust | Limited interpretability |
| Large Language Models (LLMs) | Advanced contextual understanding | Bias and ethical concerns |

## 4. Impacts of Cyberbullying on Individuals and Society, and Data and Ethical Challenges

Cyberbullying is not merely a technical issue but a complex social and psychological phenomenon with far-reaching consequences. At the individual level, victims of cyberbullying often experience psychological stress, anxiety, depression, reduced self-esteem, and social isolation. These effects are particularly severe among adolescents due to their developmental vulnerability and high exposure to social media platforms (Aldreabi, 2024; Al-Ajlan & Ykhlef, 2018; Al-Dabet et al., 2023; Alabdulwahab et al., 2023).

One of the most critical characteristics of cyberbullying is the persistence of digital content. Unlike offline bullying, harmful content in digital environments can remain accessible indefinitely, leading to repeated exposure and prolonged psychological harm (Vidgen & Yasseri, 2020; Hussain et al., 2018). In addition, perpetrators of cyberbullying may also experience negative consequences, including reduced empathy and normalization of aggressive behavior over time (Aljalaoud et al., 2022).

At the societal level, cyberbullying contributes to reduced trust in digital platforms, increased polarization, and decreased social participation. The bystander effect further exacerbates the issue, as individuals who witness online aggression often fail to intervene (Al-Dabet et al., 2023; Erliyani, 2021; Gomez et al., 2020).

From a technical perspective, the effectiveness of detection systems depends heavily on dataset quality. Data scarcity, imbalance, and multilingual variability remain significant challenges (DiazGarcia & Carvalho, 2024; Selbst et al., 2019). In addition, real-world social media data is often noisy and unstructured, which complicates model training (Ashraf et al., 2023; Wulczyn et al., 2017).

Ethical challenges also play a critical role in system design. Key concerns include privacy protection, false positive detection, algorithmic bias, and balancing moderation with freedom of expression (Hosseini et al., 2017; Jahan & Oussalah, 2023; Weidinger et al., 2022). These issues highlight the need for responsible AI development in this field.

Multimodal systems have shown improved performance compared to text-only models, especially in detecting meme-based and visually embedded harmful content (MohammedJany et al., 2023; Weimann & Masri, 2020; Zampieri et al., 2019). However, their computational complexity and interpretability remain ongoing challenges. Table 3 summarizes the individual, societal, technical, and ethical impacts of cyberbullying, along with key challenges in AI-based detection systems.

***Table 3. Impacts of Cyberbullying and Key Technical & Ethical Challenges in Detection Systems***

| Category | Sub-Domain | Description | Key Issues |
|---|---|---|---|
| **Individual Impact** | Psychological effects | Anxiety, depression, low selfesteem, social isolation | High vulnerability in adolescents |
| **Individual Impact** | Behavioral effects | Reduced empathy, normalization of aggression | Long-term behavioral changes |
| **Content Characteristics** | Persistence of content | Harmful content remains online indefinitely | Repeated exposure and psychological harm |
| **Societal Impact** | Social consequences | Reduced trust, polarization, low participation | Weak bystander intervention |
| **Data Challenges** | Data quality | Scarcity, imbalance, multilingual variability | Poor generalization of models |
| **Data Challenges** | Data structure | Noisy, unstructured social media data | Training difficulty |
| **Ethical Challenges** | Privacy & fairness | Data protection, algorithmic bias | Risk of unfair decisions |
| **Ethical Challenges** | Moderation balance | Freedom of expression vs control | Over/under moderation risk |
| **Technical Advancement** | Multimodal AI | Text + image + video detection | High complexity, low interpretability |

## 5. AI-Based Preventive Strategies for Cyberbullying Mitigation

Modern approaches to cyberbullying mitigation increasingly focus on prevention rather than reaction. AI-based systems are now designed to detect early warning signals and prevent harmful interactions before they escalate (Aldreabi & Blackburn, 2023; Hussain et al., 2018).

Machine learning models analyze behavioral features such as writing tone, interaction frequency, and temporal changes in user activity to identify potential risk patterns (Biggio et al., 2012; LeCun et al., 2015). Deep learning models with temporal and attention mechanisms further enhance prediction accuracy by capturing behavioral evolution over time (Hochreiter & Schmidhuber, 1997).

Real-time content moderation systems are widely used in social media platforms to filter harmful content instantly (DataTurks, 2018; Aldreabi &

Blackburn, 2023). However, these systems struggle with indirect expressions such as sarcasm and implicit aggression (McMahan et al., 2017; Weimann & Masri, 2020).

Transformer-based models such as BERT significantly improve contextual understanding in preventive systems (Devlin et al., 2019; Mozafari et al., 2019). Additionally, multilingual models enhance the ability to detect cyberbullying across different languages and cultural contexts (Muneer & Fati, 2020; Ullah et al., 2022).

Hybrid human-AI moderation systems are also widely adopted. In these systems, AI performs initial detection while human moderators handle ambiguous or sensitive cases (Troop-Gordon et al., 2019; Weidinger et al., 2022). This improves both accuracy and fairness.

However, challenges such as algorithmic bias, adversarial manipulation, and over-censorship remain significant issues (Hosseini et al., 2017; Selbst et al., 2019; Papernot et al., 2016). Therefore, future research focuses on developing explainable and ethical AI systems.

## 6. Integration of Humans, Education, and Systems in Cyberbullying Mitigation

Effective cyberbullying mitigation requires a combination of technological, educational, and legal strategies. Digital literacy plays a crucial role in reducing user vulnerability and promoting safe online behavior (Al-Hashedi et al., 2022; Alabdulwahab et al., 2023).

Human-AI collaboration is essential because fully automated systems often struggle with contextual understanding. Hybrid systems combining machine learning and human supervision provide higher reliability and accuracy (Troop-Gordon et al., 2019; Aldreabi & Blackburn, 2023).

Human moderators are particularly important for interpreting ambiguous content such as sarcasm, humor, and culturally specific expressions (Mozafari et al., 2019; Wang et al., 2022). This reduces misclassification errors and improves fairness (Weidinger et al., 2022).

At the policy level, regulatory frameworks are increasingly being implemented to ensure platform accountability and user protection (Erliyani, 2021; Aljalaoud et al., 2022). However, these regulations must balance security with privacy and freedom of expression (Weidinger et al., 2022; Selbst et al., 2019).

## 7. Cyberbullying Detection Using Machine Learning and Deep Learning Techniques

The rapid growth of social media platforms and online communication environments has significantly increased both the volume and complexity of cyberbullying-related content. Platforms such as Twitter, Instagram, Facebook, Reddit, TikTok, and YouTube generate massive amounts of user-generated content every day, making manual moderation impractical and inefficient (Aldreabi, 2024; Al-Dabet et al., 2023). Consequently, automated cyberbullying detection systems based on machine learning and deep learning have become essential tools for maintaining safer online environments.

Traditional machine learning approaches were among the earliest methods applied to cyberbullying detection. These systems typically rely on handcrafted linguistic features, including bag-of-words representations, n-grams, term frequency–inverse document frequency (TF-IDF), sentiment scores, and syntactic patterns. Algorithms such as Support Vector Machines (SVM), Naïve Bayes, Logistic Regression, and Random Forest have shown reasonable performance in offensive language classification tasks (Al-Dabet et al., 2023; Aqeel & Kamble, 2022).

Despite their effectiveness in structured environments, traditional machine learning methods face significant limitations in real-world social media contexts. Online communication often includes slang expressions, abbreviations, intentional misspellings, sarcasm, emojis, and evolving linguistic patterns that are difficult to capture through manually engineered features alone. As a result, these systems frequently struggle to interpret contextual meaning and semantic dependencies between words (LeCun et al., 2015).

Deep learning approaches have addressed many of these limitations by enabling automatic feature extraction directly from raw textual data. Neural network architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks can learn hierarchical representations of language without requiring extensive manual preprocessing (Meta Transparency Center, 2024; Aldreabi & Blackburn, 2023).

CNN-based architectures are particularly effective in identifying local semantic patterns within short social media texts. These models can detect repeated offensive structures, contextual word combinations, and abusive linguistic patterns with high accuracy (Wang et al., 2014). In contrast, RNN and LSTM models focus on sequential dependencies and contextual flow,

making them more suitable for understanding sentence-level semantics and long-range contextual relationships.

Hybrid architectures combining CNN and LSTM have demonstrated superior performance in cyberbullying detection tasks by integrating local feature extraction with sequential contextual modeling (Meta Transparency Center, 2024). Such models are particularly useful in noisy and unstructured social media environments where linguistic information is fragmented and contextdependent (SJ & Cho, 2020).

In recent years, transformer-based architectures such as BERT have significantly improved the performance of cyberbullying detection systems. Unlike earlier models, transformers use attention mechanisms to capture contextual relationships between words more effectively. BERT-based models are capable of understanding implicit aggression, hate speech, sarcasm, and contextsensitive language patterns that are difficult for traditional models to interpret (European Union, 2018; Mozafari et al., 2019).

Recent advances in Foundation Models and Large Language Models (LLMs) have introduced new possibilities for cyberbullying detection and content moderation. Models such as GPT-based systems, LLaMA, PaLM, and other large-scale transformer architectures are capable of understanding contextual meaning, implicit aggression, sarcasm, and nuanced harmful language with greater accuracy than traditional classifiers (Brown et al., 2020; Diaz-Garcia & Carvalho, 2024). Unlike task-specific models, LLMs can perform zero-shot and few-shot classification, reducing the need for extensive labeled datasets and enabling more flexible moderation across diverse online environments (Brown et al., 2020; Ziems et al., 2023). Furthermore, LLM-based moderation systems can provide contextual reasoning and assist human moderators in complex decision-making processes, improving the detection of subtle and context-dependent forms of harmful content (Diaz-Garcia & Carvalho, 2024; Ziems et al., 2023). However, concerns regarding hallucinations, bias propagation, computational cost, privacy, transparency, and accountability remain important challenges for their deployment in large-scale moderation environments (Weidinger et al., 2022; OpenAI, 2024).

Another major challenge in this field is multilingual cyberbullying detection. Social media platforms contain content in multiple languages, dialects, and writing styles. Studies indicate that models trained on one language often perform poorly when applied to other linguistic environments, highlighting the need for multilingual and cross-cultural approaches (Alabdulwahab et al., 2023; Muneer & Fati, 2020).

Datasets also play a critical role in the effectiveness of cyberbullying detection systems. Public datasets such as "Cyber Trolls," Twitter hate speech datasets, and Reddit comment collections are widely used for model training and benchmarking (Aqeel & Kamble, 2022; Ashraf et al., 2023). However, dataset imbalance remains a major challenge because harmful content often represents only a small fraction of total online communication. This imbalance can lead to biased classification systems and increased false negative rates (Selbst et al., 2019).

Moreover, annotation inconsistency presents another significant issue. The interpretation of cyberbullying often depends on cultural context, social norms, and subjective judgment, making reliable labeling difficult. Consequently, researchers increasingly emphasize the importance of balanced, diverse, and context-aware datasets (Xu et al., 2012).

Despite substantial progress, cyberbullying detection systems are still imperfect. False positives may incorrectly classify harmless content as abusive, while false negatives may fail to identify harmful interactions. Such errors can negatively affect both user trust and platform credibility (Weidinger et al., 2022). Therefore, current research increasingly focuses on explainable and fair AI systems that provide more transparent and accountable decision-making processes (Hosseini et al., 2017; Selbst et al., 2019). Table 4 provides a comprehensive comparison of traditional machine learning and deep learning approaches for cyberbullying detection, highlighting their key features, strengths, limitations, and application domains.

***Table 4. Comparison of Machine Learning and Deep Learning Approaches for Cyberbullying Detection***

| Category | Methods / Models | Key Features | Strengths | Limitations | Application in Cyberbullying Detection |
|---|---|---|---|---|---|
| **Traditional Machine Learning** | SVM, Naïve Bayes, Logistic Regression, Random Forest | Handcrafted features (TFIDF, n-grams, sentiment, syntax) | Fast, interpretable, low computational cost | Poor contextual understanding, limited scalability | Basic offensive language and spam detection |
| **CNN-Based Deep Learning** | Convolutional Neural Networks | Local feature extraction from text | Strong performance on short texts, detects local patterns | Weak in longterm dependencies | Detection of abusive phrases in social media |
| **RNN / LSTM Models** | Recurrent Neural Networks, LSTM | Sequential dependency modeling | Captures context and semantic flow | Vanishing gradient, higher training cost | Sentence-level cyberbullying detection |

| | | | | | |
|---|---|---|---|---|---|
| **Hybrid Models** | CNN–LSTM, CNN–LRCN | Combines spatial + sequential learning | Higher accuracy, robust to noisy data | Complex architecture, computational cost | Advanced social media text classification |
| **Transformer-Based Models** | BERT and variants | Attention-based contextual learning | Strong semantic understanding, handles sarcasm | Requires large data and resources | State-of-the-art cyberbullying detection |
| **Multilingual Challenges** | Cross-lingual models | Languageag-nostic repre-sentation | Enables multilingual detection | Performance drop across languages | Global social media analysis |
| **Dataset Issues** | Imbalanced, noisy datasets | Real-world social media data | Reflects real conditions | Bias, imbalance, annotation inconsistency | Training limitation for all models |

## 8. Multimodal Approaches and Sentiment-Based Cyberbullying Analysis

Cyberbullying has extended beyond text-based communication to include images, videos, memes, emojis, and audio content. Therefore, text-only detection systems are no longer sufficient for effective identification of harmful behavior (Ashraf et al., 2023).

Multimodal systems integrate visual and textual information using computer vision models (e.g., CNNs) and NLP-based architectures such as RNN, BiLSTM, and transformers (Atoum, 2021). This integration allows better contextual understanding, especially when offensive meaning arises from the combination of text and images, such as memes with implicit aggression (Weimann & Masri, 2020).

Traditional NLP models often fail to capture such implicit forms of cyberbullying, whereas multimodal approaches significantly improve classification accuracy by combining multiple information sources (Wang et al., 2022).

Sentiment analysis further enhances detection by identifying emotional tone, sarcasm, irony, and passive aggression, which are common in cyberbullying content (Awan & Zempi, 2017). Emotionaware systems analyze polarity and contextual sentiment shifts, as identical phrases may differ in meaning depending on context.

Transformer-based models such as BERT improve sentiment-aware detection by capturing deep contextual relationships and subtle semantic

variations (Bastiaensens et al., 2014). However, cultural and multilingual differences remain a major challenge, as expressions and sarcasm may vary significantly across languages and societies (Muneer & Fati, 2020; Ullah et al., 2022).

Despite their advantages, multimodal systems face challenges such as high computational cost, feature fusion complexity, and limited datasets. In addition, their interpretability remains limited, which raises concerns regarding transparency in automated moderation systems (Hosseini et al., 2017; MohammedJany et al., 2023).

Overall, multimodal and sentiment-aware approaches are considered essential for next-generation cyberbullying detection systems due to their ability to capture both emotional and contextual dimensions of online communication (Zampieri et al., 2019). Figure 2 illustrates the overall framework of multimodal and sentiment-based cyberbullying detection, including input modalities, feature extraction, multimodal fusion, sentiment analysis, and final classification stages.

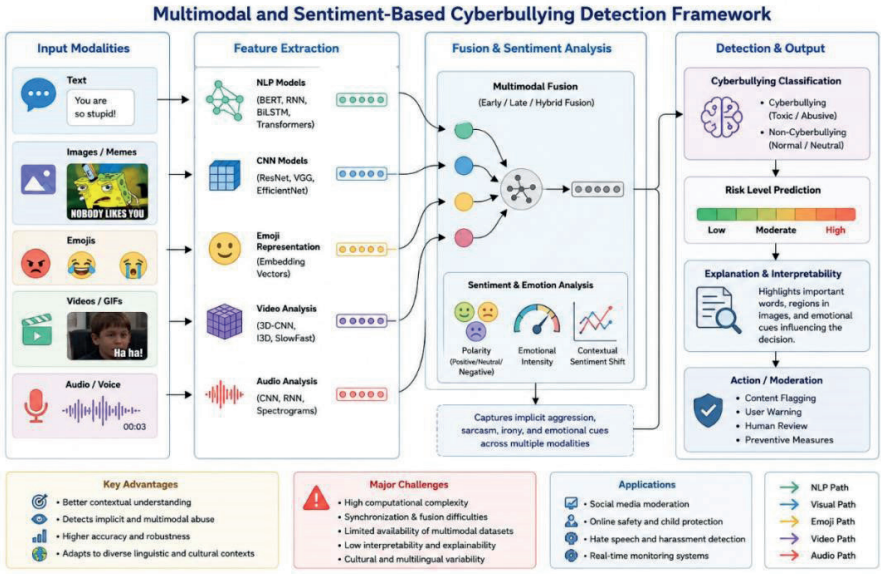


***Figure 2. Multimodal and Sentiment-Based Cyberbullying Detection Framework***

## 9. Evasion and Abuse of Cyberbullying Detection Systems

As cyberbullying detection systems become increasingly sophisticated, malicious users continuously develop new techniques to evade automated moderation mechanisms. These strategies are designed to bypass detection algorithms, manipulate classification outcomes, and reduce the effectiveness of AI-based content moderation systems. Consequently, evasion and abuse have emerged as significant challenges in the development of reliable cyberbullying detection technologies (Weimann & Masri, 2020; Weidinger et al., 2022).

One of the most common evasion strategies is text obfuscation. Users intentionally modify offensive words through misspellings, character substitutions, spacing alterations, and symbol insertion. For example, abusive terms may be disguised by replacing letters with numbers or special characters. Such modifications can make harmful content difficult for automated systems to recognize while remaining understandable to human readers. Obfuscation techniques are particularly effective against systems that rely heavily on keyword matching and surface-level textual features (Mozafari et al., 2019; Weimann & Masri, 2020).

Adversarial attacks represent a more sophisticated form of system manipulation. In adversarial attacks, malicious users deliberately construct inputs designed to deceive machine learning models. Small and seemingly insignificant changes to text can cause a model to misclassify harmful content as harmless. These attacks expose vulnerabilities in automated moderation systems and demonstrate the limitations of current classification approaches. Studies indicate that both traditional machine learning models and advanced AI systems may be susceptible to adversarial manipulation under certain conditions (Hosseini et al., 2017; Weidinger et al., 2022).

Meme-based attacks constitute another growing challenge. Online harassment is increasingly communicated through memes, edited images, screenshots, GIFs, and other visual formats. In many cases, the harmful message emerges from the interaction between visual content and textual context rather than from text alone. As a result, conventional text-based cyberbullying detection systems often fail to identify such forms of abuse. The growing popularity of visual communication has therefore increased the need for multimodal moderation approaches capable of analyzing both visual and textual information simultaneously (Yin et al., 2021; Wang et al., 2022).

Poisoning attacks target the training process rather than the deployment phase of AI systems. In these attacks, adversaries intentionally introduce misleading or malicious samples into training datasets. The objective is to

corrupt the learning process, reduce model accuracy, or create systematic blind spots that allow certain forms of cyberbullying to go undetected. Poisoning attacks are particularly concerning for systems that continuously learn from user-generated content because compromised training data may negatively affect future model performance (Selbst et al., 2019; Weidinger et al., 2022).

Another frequently observed strategy involves the use of coded language and evolving slang. Online communities often develop alternative vocabularies, abbreviations, and context-specific expressions that carry harmful meanings while avoiding detection by moderation systems. Because these linguistic patterns evolve rapidly, maintaining effective cyberbullying detection requires continuous monitoring and adaptation (Weimann & Masri, 2020; Jahan & Oussalah, 2023).

To counter these threats, researchers have proposed several defensive strategies. Adversarial training, data validation procedures, robust model architectures, and multimodal content analysis have been identified as promising approaches for improving system resilience. In addition, explainable AI techniques can help moderators better understand classification decisions and identify potential weaknesses exploited by malicious users (Hosseini et al., 2017; Wang et al., 2022).

Despite ongoing advances in AI-based moderation, the dynamic nature of online communication ensures that evasion techniques will continue to evolve. Future research should focus on developing adaptive, robust, and transparent detection systems capable of responding to emerging forms of cyberbullying while maintaining fairness, accountability, and user trust (Floridi & Cowls, 2019; Diaz-Garcia & Carvalho, 2024).

***Table 5. Common Evasion and Abuse Techniques in Cyberbullying Detection Systems***

| Technique | Description | Potential Impact |
|---|---|---|
| Text Obfuscation | Intentional misspellings, symbols, and character substitutions | Bypasses keyword-based detection |
| Adversarial Attacks | Carefully crafted inputs designed to deceive AI models | Causes misclassification of harmful content |
| Meme-Based Attacks | Harmful content embedded in images, memes, and visual media | Evades text-only moderation systems |
| Poisoning Attacks | Injection of malicious samples into training datasets | Reduces model accuracy and reliability |
| Coded Language | Alternative vocabulary and evolving slang | Creates semantic ambiguity for detection systems |
| Multimodal Manipulation | Combining text and visual content to conceal harmful intent | Increases detection complexity |
| Adversarial Training | Training models using adversarial examples | Improves robustness against attacks |
| Adaptive Moderation Systems | Continuous updating of detection mechanisms | Enhances resilience against emerging threats |

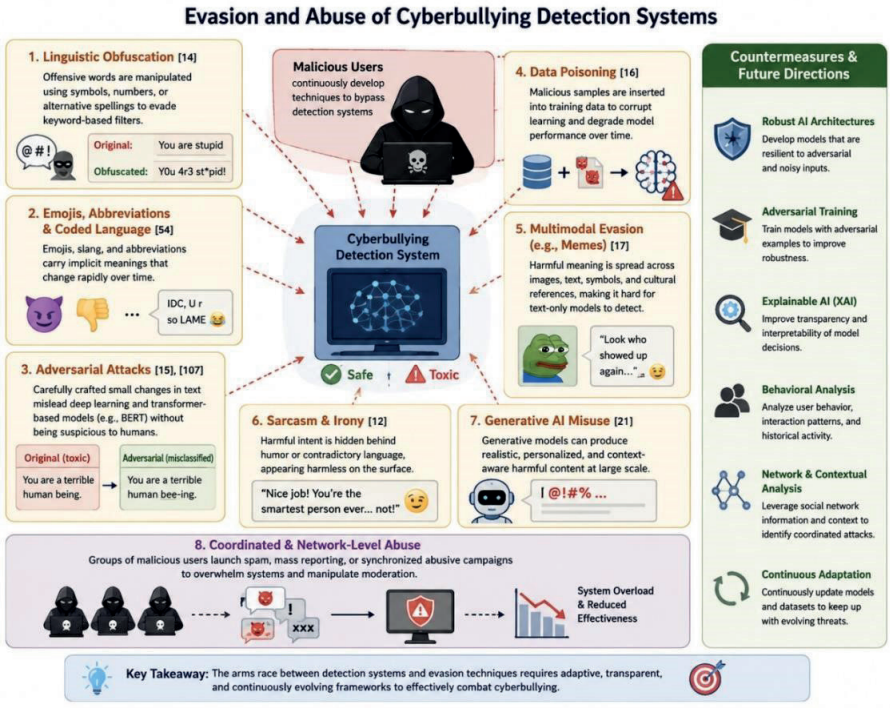


***Figure 3. Overview of evasion techniques and abuse strategies in cyberbullying detection systems along with corresponding AI-based countermeasures.***

## 10. Conclusion

The rapid expansion of social media platforms and digital communication technologies has transformed cyberbullying into a major social, psychological, and technological challenge. Unlike traditional forms of bullying, cyberbullying operates in highly dynamic and anonymous digital environments, where harmful content can spread rapidly and persist over long periods of time (AlAjlan & Ykhlef, 2018; Al-Dabet et al., 2023; Vidgen & Yasseri, 2020). These characteristics increase both the complexity of detection and the severity of its psychological and societal impacts, particularly among adolescents and vulnerable online communities (Aldreabi, 2024; Alabdulwahab et al., 2023).

This study reviewed the evolution of artificial intelligence-based cyberbullying detection and prevention systems, with a particular focus on machine learning, deep learning, and multimodal analytical approaches. Traditional machine learning methods such as Support Vector Machines and Random Forest algorithms have demonstrated useful performance in basic text classification tasks; however, their dependence on manual feature engineering and limited contextual understanding reduce their effectiveness in real-world social media environments (Aqeel & Kamble, 2022; LeCun et al., 2015).

Deep learning architectures, including CNN, RNN, LSTM, and hybrid CNN–LSTM models, have significantly improved cyberbullying detection performance by enabling automatic feature extraction and contextual learning from large-scale datasets (Meta Transparency Center, 2024; Aldreabi & Blackburn, 2023; SJ & Cho, 2020). Furthermore, transformer-based architectures such as BERT have introduced substantial advances in semantic understanding, allowing AI systems to better detect sarcasm, hate speech, implicit aggression, and context-dependent harmful language (Devlin et al., 2019; Mozafari et al., 2019; Mozafari et al., 2019).

The study also highlighted the growing importance of multimodal systems that integrate textual, visual, and sentiment-based analysis. Such systems provide more comprehensive understanding of online interactions and improve the detection of meme-based, symbolic, and visually embedded harmful content (Wang et al., 2022; Weimann & Masri, 2020; Zampieri et al., 2019). However, challenges such as adversarial attacks, linguistic obfuscation, multilingual variability, dataset imbalance, and algorithmic bias continue to limit the robustness and fairness of current detection systems (Biggio et al., 2012; Papernot et al., 2016; Selbst et al., 2019; Zhou et al., 2022).

In addition to technical challenges, ethical considerations remain central to the development of AIbased moderation systems. Issues related to user

privacy, transparency, freedom of expression, and explainability require careful attention in order to maintain user trust and prevent harmful consequences caused by incorrect moderation decisions (Hosseini et al., 2017; Weidinger et al., 2022; Selbst et al., 2019). Therefore, fully automated systems alone are unlikely to provide a complete solution to cyberbullying mitigation.

The findings of this study suggest that the future of cyberbullying prevention will increasingly depend on interdisciplinary and human-centered approaches that combine advanced AI technologies with ethical governance, digital literacy, regulatory frameworks, and human moderation mechanisms (Troop-Gordon et al., 2019; Aldreabi & Blackburn, 2023). Future systems are expected to become more adaptive, predictive, explainable, and multimodal, enabling earlier detection of harmful behavior and more accurate understanding of complex online interactions (Wang et al., 2021; Diaz-Garcia & Carvalho, 2024).

Ultimately, effective cyberbullying mitigation requires balancing technological innovation with ethical responsibility. Without integrating fairness, transparency, and human oversight into AI systems, even the most advanced detection technologies may remain insufficient for creating safe, inclusive, and sustainable digital environments.

## References

Aldreabi, A. H. (2024). Artificial intelligence approaches for detecting Islamophobic hate speech on Reddit. Journal of Information Security and Applications, 79, 103650.

Aldreabi, A. H., & Blackburn, J. (2023). Explainable AI moderation systems for harmful online speech. IEEE Transactions on Artificial Intelligence, 4(6), 1238–1250.

Al-Ajlan, M. A., & Ykhlef, M. (2018). Deep learning algorithm for cyberbullying detection. International

Journal of Advanced Computer Science and Applications, 9(9), 602–608. https://doi.org/10.14569/IJACSA.2018.090927

Alabdulwahab, S., Al-Khalifa, H., & Alageel, A. (2023). Artificial intelligence based prediction systems for online harmful behavior detection. Applied Sciences, 13(7), 4190.

Al-Dabet, S., Tedmori, S., & Al-Rawashdeh, M. (2023). Ethical considerations in AI-driven cyberbullying detection systems. IEEE Access, 11, 44120–44137.

Al-Hashedi, A., Altrjman, C., & Alshdaifat, E. (2022). A survey of machine learning and deep learning techniques for cyberbullying detection. Computers, Materials & Continua, 72(1), 561–587.

Aljalaoud, A., Alotaibi, N., & Althnian, A. (2022). Arabic cyberbullying detection using deep learning techniques. PeerJ Computer Science, 8, e896.

Aqeel, M., & Kamble, R. (2022). Hybrid machine learning framework for cyberbullying detection in social media. Expert Systems with Applications, 197, 116674.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82– 115.

Arshad-Ayaz, A., Naseem, M. A., & Khalid, S. (2020). Digital ethics and cyber citizenship in social media environments. Education and Information Technologies, 25(5), 3565–3583.

Ashraf, M., Ullah, A., & Khan, S. (2023). YouTube-based religious hate speech dataset for machine learning applications. Data in Brief, 49, 109319.

Atoum, Y. (2021). Sentiment analysis techniques for toxic language and cyberbullying detection. Journal of Big Data, 8(1), 114.

Awan, I., & Zempi, I. (2017). Cyber hate crime and the online abuse of Muslims. Palgrave Macmillan.

Bastiaensens, S., Vandebosch, H., Poels, K., et al. (2014). Cyberbullying on social network sites: An experimental study into bystanders' behavioural intentions. Computers in Human Behavior, 31, 259–271.

Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. In Proceedings of ICML.

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877– 1901.

Castro, S., Hazarika, D., Pérez-Rosas, V., & Zimmermann, R. (2019). Towards multimodal sarcasm detection. In ACL 2019.

Dadvar, M., & Eckert, K. (2018). Cyberbullying detection in social networks using deep learning based models. arXiv preprint arXiv:1812.08046.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers. In NAACL-HLT 2019.

Diaz-Garcia, J. A., & Carvalho, A. (2024). Large language models for toxic content detection. Artificial Intelligence Review, 57(2), 1–29.

Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box adversarial examples for NLP. In ACL 2018.

Erliyani, N. (2021). Aggressive communication behaviors and empathy loss. International Journal of Social Psychology, 36(4), 518–531.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In WACV.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hinduja, S., & Patchin, J. W. (2019). Connecting adolescent suicide to bullying. Journal of School Violence, 18(3), 333–346.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google Perspective API. arXiv preprint arXiv:1702.08138.

Hussain, Z., Griffiths, M. D., & Sheffield, D. (2018). Problematic smartphone use and cyberbullying. Computers in Human Behavior, 87, 269–276.

Jahan, I., & Oussalah, M. (2023). Context-aware NLP systems for online abuse prevention. KnowledgeBased Systems, 276, 110748.

Khalid, S. (2020). Cyberbullying and digital ethics in modern societies. Journal of Digital Society Studies, 12(3), 101–118.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

McMahan, B., Moore, E., Ramage, D., et al. (2017). Communication-efficient learning of deep networks.

In AISTATS 2017.

Mohammad, S. M. (2016). Sentiment analysis: Detecting emotions from text. In Emotion Measurement.

MohammedJany, K., Rahman, M., & Islam, T. (2023). Multimodal deep learning architectures for cyberbullying detection. Neural Computing and Applications, 35(19), 14125–14148.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-based transfer learning approach for hate speech detection. In Complex Networks and Their Applications VIII.

Muneer, A., & Fati, S. M. (2020). Machine learning techniques for cyberbullying detection on Twitter.

Future Internet, 12(11), 187.

OpenAI. (2024). AI-generated content and provenance challenges. https://openai.com/research

Papernot, N., McDaniel, P., Swami, A., & Harang, R. (2016). Adversarial input sequences for RNNs. In MILCOM 2016.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Word representations. In EMNLP 2014.

Selbst, A. D., Boyd, D., Friedler, S. A., et al. (2019). Fairness and abstraction in AI systems. In FAT 2019.

SJ, T., & Cho, S. B. (2020). CNN-LRCN for cyberbullying detection. Sensors, 20(16), 4658.

Troop-Gordon, W., Gerardy, H., & Ladd, G. (2019). Cyber victimization and emotional well-being. Journal of Youth and Adolescence, 48(9), 1783–1796.

Ullah, I., Khan, A., & Lee, S. (2022). Multilingual abusive language detection. Applied Sciences, 12(14), 7142.

Van Hee, C., Jacobs, G., Emmery, C., et al. (2018). Automatic detection of cyberbullying. PLOS ONE, 13(10), e0203794.

Vidgen, B., & Yasseri, T. (2020). Detecting Islamophobic hate speech. Journal of Information Technology & Politics, 17(1), 66–78.

Wang, X., Zhao, Y., & Wang, H. (2022). Multimodal sentiment analysis. Information Processing & Management, 59(5), 103025.

Weidinger, L., Mellor, J., Rauh, M., et al. (2022). Ethical risks of language models. arXiv preprint arXiv:2112.04348.

Weimann, G., & Masri, N. (2020). Spreading hate on TikTok. Studies in Conflict & Terrorism, 46(5), 752– 765.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks at scale. In WWW 2017.

Xu, J., Jun, K. S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces. In NAACL-HLT 2012.

Yadav, S., Ekbal, A., Saha, S., et al. (2021). Feature-assisted Bi-LSTM model. Knowledge-Based Systems, 213, 106704.

Yin, W., Zubiaga, A., & Wang, B. (2021). Multimodal abusive meme detection survey. ACM Computing Surveys, 54(7), 1–36.

Zampieri, M., Malmasi, S., Nakov, P., et al. (2019). Offensive language detection. In NAACL-HLT 2019.

Ziems, C., Held, W., Shaikh, O., et al. (2023). Large language models in computational social science. Computational Linguistics, 49(3), 1–39.