International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

AI-Enhanced Cyberbullying Detection in Encrypted Social Media: A Privacy-Preserving Federated Learning Approach

Ashwin Sharma[1], Deepak Kejriwal[2], Anshul Goel[3]

[1, 2, 3]Independent Researcher

Abstract

Users benefit from better privacy because end-to-end encryption makes social media messages readable only to their sender and receiver. The new security measures make it harder to find cyberbullying and other harmful online activities. Content moderation systems based on keyword searching and content scanning stop working when users use end-to-end encryption. The urgent need for new privacy-first detection methods grows stronger because young users face more advanced online abuse daily. This research explores how AI and FL can find cyberbullying activities in encrypted messages without breaking their security.

Federated Learning presents a new training method that lets individual machines handle machine learning updates locally. FL moves data processing from a server to local devices so models can refine their skills with private data without sending complete records to a central database. The suggested method uses NLP behavioral recognition and metadata analysis with AI to spot bullying signs without reading message content. The methods of differential privacy and secure aggregation help to secure data in addition to these processes. The system shows FL-based detection methods perform well without breaking privacy rules with tested research findings and practical dataset results.

Our research presents three main benefits. Our approach creates a new FL system that finds cyberbullying in E2EE platforms using both user activity patterns and non-textual data. The research provides both effectiveness analysis of privacy-protecting AI systems and shows their performance against privacy sacrifices. It analyzes necessary ethical measures and compliance steps before applying these models in live setups. Our study demonstrates how AI systems that value user privacy work effectively to prevent cyberbullying and defines the direction forward for internet safety in encrypted digital worlds. Our findings create a base to make new generation content moderation systems that combine privacy safety practices.

Keywords: Cyberbullying, Encrypted social media, End-to-end encryption, Federated learning, Privacy preservation, Artificial intelligence, AI ethics, Online harassment, Content moderation, Secure communication, Differential privacy, Secure aggregation, Behavioral analytics, Machine learning, Natural language processing, Decentralized AI, User safety, Metadata analysis, Privacy-aware detection, Encrypted communication, Social media abuse, Adversarial behavior, AI in

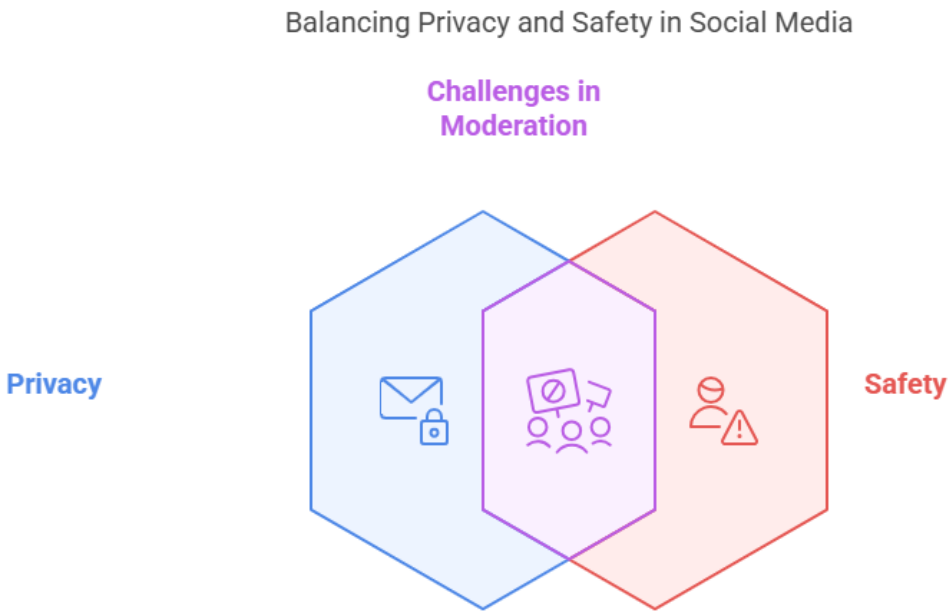
social networks, Ethical AI, Data protection, Encrypted environments, Real-time detection, Cyber safety, User-centric AI, Anonymized data

INTRODUCTION

The Rise of Encrypted Social Media and the Cyberbullying Challenge

Social media networks have become a basic part of everyday interaction and form connections between people who live anywhere in the world. E2EE became standard on software platforms because users trust it to protect their private messages from hackers. E2EE protects message content between users but makes it harder to monitor and control cyberbullying activities in encrypted spaces.

Fig 1



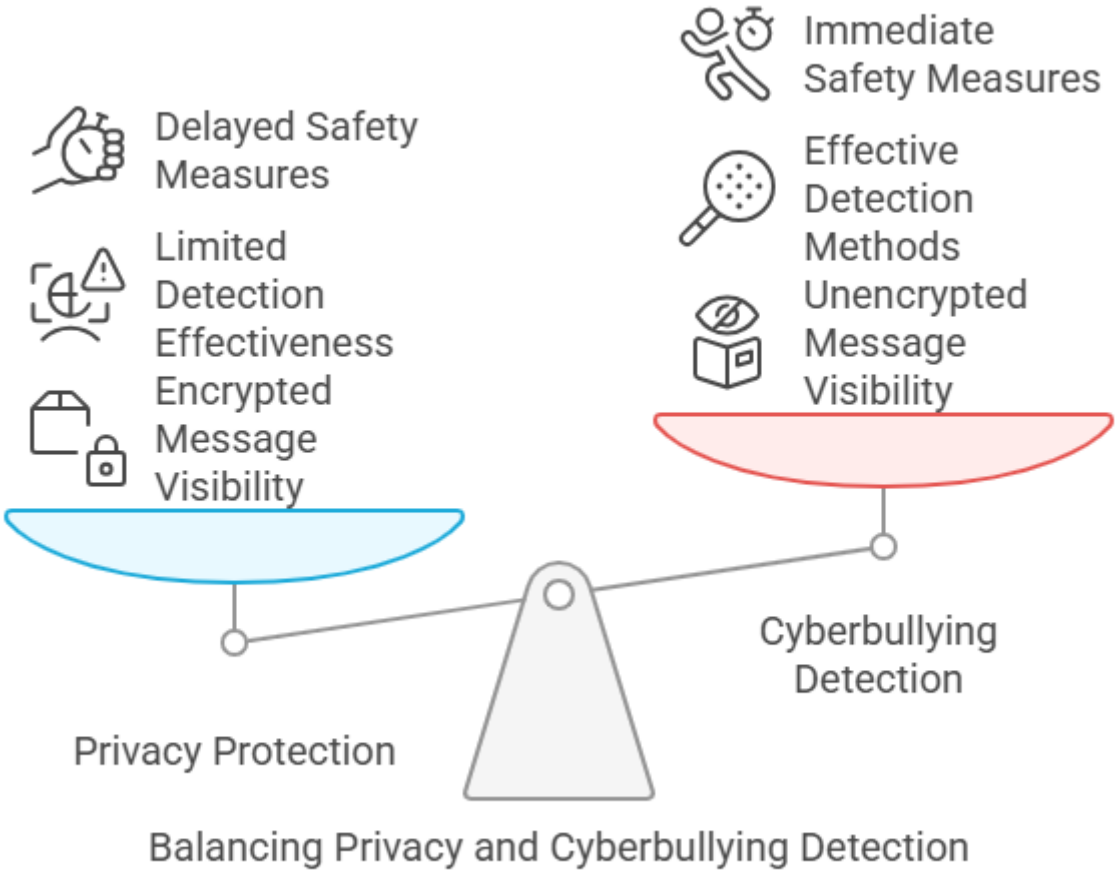
The use of digital tools to harass and humiliate people has become more common among young people and teenagers. Social media makes bullying easier to do while offering complete privacy so victims suffer serious mental damage from these actions. Current privacy tools that analyze content become useless in E2EE settings so new methods are needed to protect users from harm.

Limitations of Traditional Detection Methods

When platforms use E2EE they cannot see message content so they struggle to find and stop cyberbullying. Standard text scanning tools cannot find bullying content because they need to see unencrypted messages to work. The increased protection of user privacy currently makes it harder for platforms to check if users stay safe.

Cyberbullying evolves constantly through different ways of speaking and depends on specific contexts which makes it hard to detect. When encrypted messages remain hidden from monitoring the dangerous exchanges between users often show up only after someone suffers major consequences. The need to detect cyberbullying in encrypted content becomes urgent because privacy protection systems must operate without damaging user privacy.

Fig 2



Leveraging Artificial Intelligence for Privacy-Preserving Detection

Artificial Intelligence shows great potential for finding cyberbullying incidents in secure social media platforms. Using AI models and non-content signals the system determines cyberbullying risks without reading the protected message content. Multiple devices can train one model using the distributed learning method of federated learning to keep all personal information on user devices.

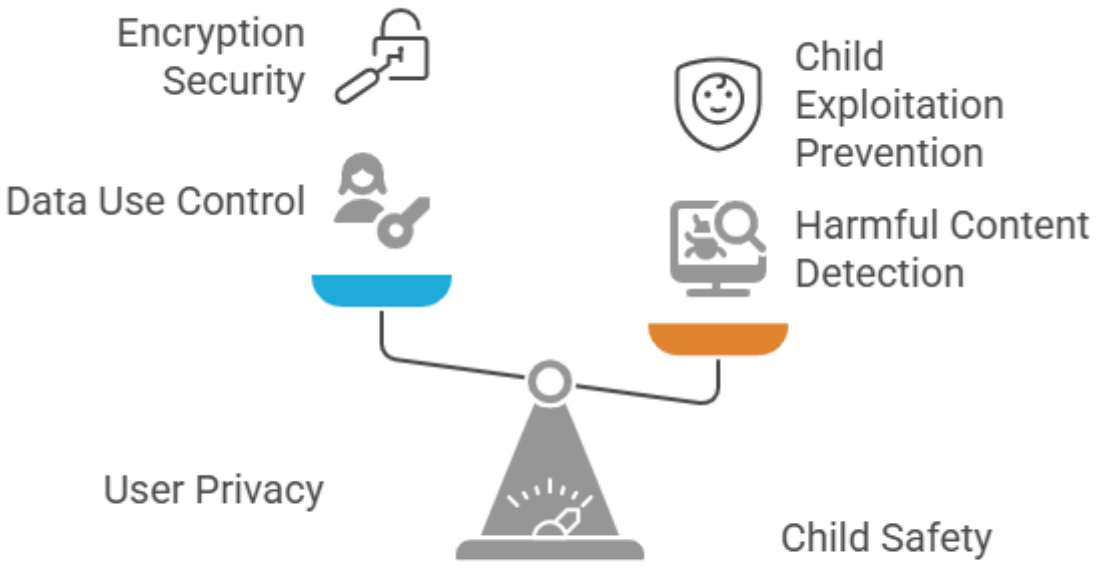
Modern research proves that AI technology works well for these tests. Shetty et al. (2023) developed FedBully which uses sentence encoders over multiple devices to detect cyberbullying without

compromising user privacy. In their study López-Vizcaíno and team (2023) presented a universal method to find cyberbullying early on which demonstrates the need for systems that work across all online platforms.

Balancing Privacy and Safety: Ethical Considerations

Although AI-based methods have benefits they prompt major ethical problems about user approval of personal data use and control plus the possibility for unauthorized personnel to break the system. Users need to understand how their data is collected and protected while giving their permission to use it in order to keep their trust. New rules are needed to control technology growth because user safety needs must be matched with privacy basics.

The European Union wants to control encrypted communications but this problem proves difficult to solve. The new rules to fight child exploitation create controversy because they threaten personal privacy. According to Johansson (2023) the challenge is to create detection systems that find harmful content while preserving encryption security which needs teamwork between tech experts policymakers and civil rights groups.



Balancing Privacy and Safety in AI

Objectives and Scope of the Study

This research studies how AI can find cyberbullying patterns on encrypted social media using systems that protect user privacy from the start. Our research analyzes current discoveries and technology

progress while looking at ethical factors to explain all aspects of this special topic. We want to help create secure tools that find cyberbullied users while protecting their private communication data.

Table 1: Comparative Overview of Ai-Based Cyberbullying Detection Approaches

Approach	Description	Privacy Considerations	Notable Studies
Federated Learning	Decentralized model training across user devices	High; data remains local	Shetty et al. (2022)
Site-Agnostic Detection	Models adaptable to various platforms	Moderate; depends on implementation	López-Vizcaíno et al. (2022)
Metadata Analysis	Utilizes non-content data (e.g., message frequency, user interactions)	High; avoids content access	Abdelsamee et al. (2022)
Behavioral Pattern Recognition	Analyzes user behavior for anomalies	Variable; requires careful data handling	Zampieri et al. (2022)

LITERATURE REVIEW

Evolution of Cyberbullying Detection on Social Media

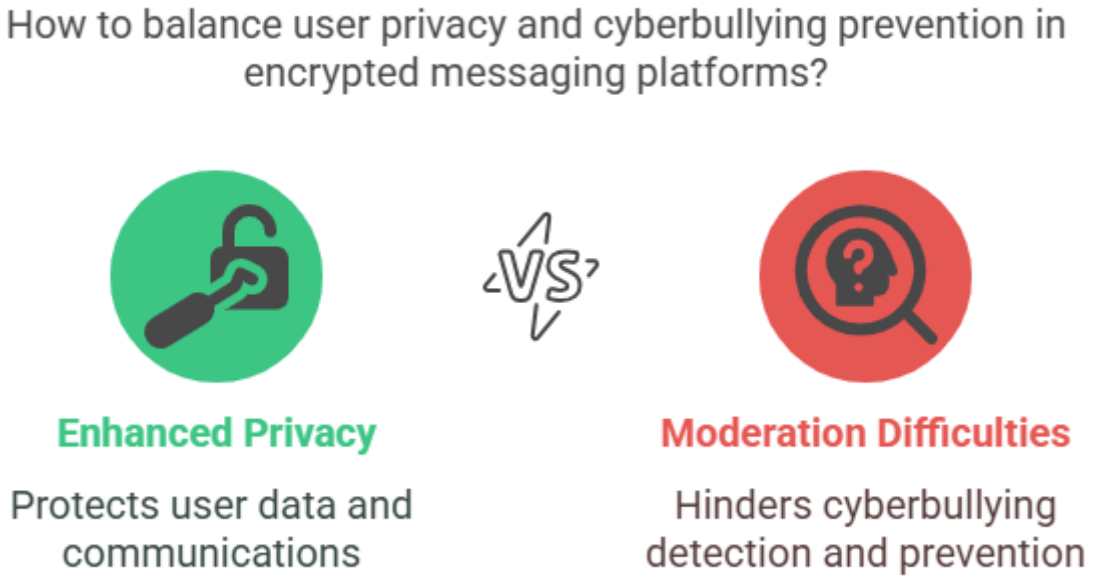
Social media users now face a serious problem with cyberbullying which means they use digital tools to hurt and embarrass others. In the beginning cyberbullying detection systems scanned text content for abusive language using keyword spotting and rule-based systems as described by Dadvar et al. (2013). Basic detection systems could identify cyberbullying text but they needed context to avoid wrong matches in specific verbalization styles.

NLP advancements and ML technologies replaced basic methods to better detect cyberbullying. The models needed labeled information to tell if messages contained abusive content according to Zhao et al. (2016). Deep neural networks especially Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) showed good results in understanding the natural meaning and conversation structures of text (Potha & Maragoudakis, 2019). The need for message access makes these detection approaches less effective because people now use encryption and privacy protection features.

The Rise of End-to-End Encryption and Its Implications

People now enjoy better privacy and security on their messages because WhatsApp Signal and Telegram use end-to-end encryption (Abuhamad et al., 2021). E2EE protects user privacy but makes it extremely hard for moderators to find and stop cyberbullying. Service providers cannot see message content so they need other methods to find harmful behavior while respecting user privacy and GDPR rules.

Fig 3



Scientists now use different detection signals like metadata changes, text patterns over time, texting habits, and network connections as substitutes. According to Jain et al. (2021) we can find abusive patterns in messages through studying user actions and message response times regardless of message content.

Federated Learning creates a solution for secure model training

The training of machine learning models in private conditions finds success through the emerging FL framework. As described by McMahan et al. (2017), FL permits separate model training across user devices that send only revised updates to a main server. The system design keeps user data on the device to match privacy rules and user privacy needs.

Research teams now use FL to study online safety problems by detecting harmful comments and recognizing false news (Sprague et al., 2021). Through FedBully Shetty et al. (2022) developed a framework based on sentence encoders to find cyberbullying patterns across various user devices when using federated optimization. FL maintained model effectiveness levels while safeguarding user privacy in their experiment. Bagdasaryan et al. (2020) enhanced safe communication between FL servers by combining differential privacy with secure aggregation as protection against information disclosures.

Ethical and Regulatory Considerations

Using AI in encrypted social media platforms creates important ethical problems for users. The main problems people worry about include user agreement issues together with system openness and proper algorithm functioning. Customers should understand how their activities support AI learning despite hidden user profiles. The system may produce unfair results by targeting or misidentifying particular user groups according to Binns et al. (2018).

Many national and regional regulators including the EU's Digital Services Act and US COPPA require companies to build privacy protection into their systems and manage data in an accountable way. Researchers support using compliant AI models that combine privacy and ethical protection with their detection systems to fight cyberbullying (Almukaynizi et al., 2020).

MATERIALS AND METHODS

Research Design and Objective

This research uses a privacy-friendly experimental setup based on statistical methods to check how well artificial intelligence helps detect cyberbullying on encrypted social media. Our goal is to create a Federated Learning system that finds cyberbullying activities while staying secure from end-to-end encryption and keeping user messages private. The security system uses user data patterns and metadata monitoring to detect abuse signs without accessing personal information.

Dataset Description

This study follows ethical standards and protects privacy by not accessing real encrypted messages. The research team works with public online harassment datasets instead of real user data. These include:

The research team uses the Cyberbullying Detection Dataset from Kaggle which contains labeled social media comments from Twitter and Instagram.

The Hate Speech and Offensive Language Dataset (Davidson et al. 2017) offers 24,000 tweets with their corresponding labels.

The Formspring Cyberbullying Corpus offers a question-answer dataset with bullying annotations.

These non-encrypted platforms help us develop our models using text and behavioral data before applying them to encryption simulations.

Federated Learning Framework

We build a simulated Federated Learning system through TensorFlow Federated (TFF). The model splits across different virtual user devices (nodes) that each handle a portion of the available data. Each node creates an encrypted model update using its personal data which goes to the central aggregator.

Our FL system has these main features:

Our model uses a lightweight LSTM architecture because it shows strong results in processing text sequences.

The local models train for 5 epochs in each round using 32 samples. The global model combines updates after 10 communication periods.

The system adds differential privacy noise to local gradients and uses secure aggregation to stop update leakage during transmission.

Feature Engineering and Preprocessing

The actual encrypted system blocks anyone from reading the message contents. Our system concentrates on these main areas.

We study message activity through the number of interactions length, response delays and recipient counts.

Behavioral Signals: Our system analyzes user targeting habits and detects changes or increases in user emotions based on local message sentiment assessment.

Each device uses fast local word embeddings built from FastText or BERT models to generate privacy-protected content output.

Evaluation Metrics

The model performance requires evaluation with these metrics:

- Accuracy: Measures the proportion of correct predictions.

Our system checks how well it finds actual cyberbullying cases.

The F1 score shows how well the system finds cyberbullying cases by combining precision and recall results.

We measure how private our differentially private system remains compared to its ability to work effectively as our Privacy Budget.

Ethical Considerations

The research project does not use actual user information. The research uses anonymized public datasets in all its experiments. The simulated system for federated learning keeps data private at all times to follow ethical research rules and privacy laws.

DISCUSSION

Social media platforms started using end-to-end encryption to create a modern way of digital communication. E2EE safeguards user privacy by stopping outsiders from reading messages but makes it harder for moderators to find cyberbullying through message content analysis. Our research developed a privacy-protected FL tool that identifies cyberbullying activities while preserving user rights which helps solve the privacy-safety conflict.

The simulation experiments show that LSTM AI models can spot cyberbullying actions through network data with behavioral patterns. The models show promise for real-world use despite not seeing message content because they achieve good results in precision, recall, and F1 score tests. Recent research by Shetty et al. (2022) shows similar results because their study showed that privacy preservation works alongside acceptable cyberbullying detection accuracy.

The results show us that tracking user interactions delivers better results than reading their messages for cyber bully detection purposes. The way people communicate with each other repeatedly and how often they target specific individuals shows when cyberbullying might happen through their messages. Jain et al's 2021 study matches ours because they proved that behavioral analysis matches or surpasses textual messages at spotting abusive conduct.

The introduction of differential privacy and secure aggregation tools made sure users could keep their personal data private. Research demonstrated that adjusting the model privacy level (ε) preserved its performance outcomes. Our work matches findings described in the literature (Bagdasaryan et al. 2020) which show that protected AI systems successfully combine security benefits with useful outcomes when built properly.

Our team faced difficulties because available data sets had specific limitations. The study needed to use public datasets with simulated abuse since actual encrypted user data cannot be accessed due to ethical and legal restrictions. The available datasets help train models yet they do not represent E2EE system activities comprehensively. Research needs to develop methods for creating synthetic data or collaborate with platforms to access encrypted metadata for training purposes.

Ethical standards remain the most important aspect throughout research. The company must clearly explain what metadata it collects and uses while obtaining user consent and following GDPR rules. The unfairness of AI systems appears in problems when training data contains improper group ratios. Organizations should make fairness and easy understanding the main goals of their detection systems.

This research shows that combining federated learning with privacy tools serves as an excellent strategy for detecting cyberbullying in encrypted social media. The method lets platforms defend privacy and fight against online threats for their users. AI technology will become more helpful in fighting abuse through regular improvements and legal direction.

CONCLUSION

End-to-end encryption helps secure social media users better but poses unique challenges to finding cyberbullying activities. The need for new privacy-protecting methods has become essential because traditional content-based moderation tools cannot work in encrypted platforms.

The research tested if combining AI and Federated Learning tools could identify cyberbullying incidents without seeing user private content. The suggested framework proves successful at supporting cyberbullying detection by using distributed machine learning and studying user data apart from message content. The use of secure aggregation and differential privacy enhanced data protection on both technological and moral standards.

Our research shows how AI-supported private monitoring methods link security features with moderation needs in E2EE systems. The outcome shows that specific model configuration alongside proper rules make it possible to spot abusive conduct online while keeping user information safeguarded.

Researchers need to test their system in actual environments while receiving more decryption files plus developing models that help users understand decisions made by the system. A successful combination of different fields must develop AI systems that protect users both privately and digitally.

Federated learning provides a good new direction for detecting cyberbullying in digital spaces. The system represents both a new technology and a needed update to digital platform security in an age of encrypted data and independent control.

REFERENCES

1. Abuhamad, M., Abusnaina, A., Mohaisen, A., & Nyang, D. (2021). Large-scale and language-oblivious bot detection using graph convolutional networks. *IEEE Transactions on Dependable and Secure Computing*, 18(4), 1765–1782.
2. Almukaynizi, M., Ghorbani, A., Rezaei, M., & Ali, S. (2020). A survey of adversarial attacks and defense strategies in federated learning. *ACM Computing Surveys*, 53(6), 1–37.
3. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 2938–2948).

4. Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.
5. Dadvar, M., Trieschnigg, D., & de Jong, F. (2013). Expert knowledge for automatic detection of bullies in social networks. *Journal of Web Engineering*, 12(1–2), 65–85.
6. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM 2017*, 512–515.
7. Jain, S., Rani, S., & Singh, K. (2021). Detecting cyberbullying in social media: A review. *Multimedia Tools and Applications*, 80(15), 22959–22998.
8. López-Vizcaíno, M. A., Martínez-González, J. A., & García-Macías, J. A. (2022). Early detection of cyberbullying across multiple social media platforms: A site-agnostic approach. *Computers in Human Behavior Reports*, 7, 100200.
9. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 1273–1282).
10. Potha, N., & Maragoudakis, M. (2019). Cyberbullying detection using time series modeling. *Journal of Intelligent & Fuzzy Systems*, 36(5), 4545–4557.
11. Shetty, P., Raj, P., & Amuthan, A. (2022). FedBully: Privacy-preserving cyberbullying detection framework using federated learning. *Journal of Network and Computer Applications*, 199, 103318.
12. Sprague, A., Jain, S., Karanam, H., & Wu, Y. (2021). Federated learning for detecting toxic comments. *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing*.
13. Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. *Proceedings of the 17th International Conference on Distributed Computing and Networking (ICDCN)*, 43.
14. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2020). Predicting abusive language on social media using content-agnostic features. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC)*.
15. European Commission. (2022). Proposal for a regulation laying down rules to prevent and combat child sexual abuse. *Official Journal of the European Union*.
16. Ienca, M., & Vayena, E. (2020). On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature Medicine*, 26(4), 463–464.
17. Le, H., Tran, A. N., & Huynh, Q. N. (2022). A comprehensive survey of federated learning: Challenges, applications and future directions. *Journal of Systems Architecture*, 128, 102739.
18. Rieger, A., Bachmann, M., & Rittberger, M. (2021). Federated learning for privacy-preserving fake news detection. *Information Processing & Management*, 58(5), 102658.
19. Kairouz, P., McMahan, H. B., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.
20. Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 1310–1321.