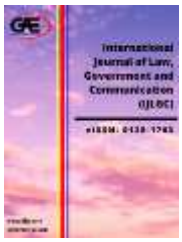
International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440



INTERNATIONAL JOURNAL OF LAW, GOVERNMENT AND COMMUNICATION (IJLGC)
www.gaexcellence.com/ijlgc

A REVIEW OF CYBERBULLYING DETECTION ON SOCIAL MEDIA

Mohd Azizuddin Ali [1*], Noor Suhana Sulaiman [2], Ziti Fariha Mohd Apandi [3]

[1]Department of Computer Network, University College TATI, Malaysia
azizs@uctati.edu.my https://orcid.org/0009-0002-3385-5585
[2]Department of Computer Network, University College TATI, Malaysia
suhana@uctati.edu.my https://orcid.org/0000-0003-1041-5296
[3]Department of Computer Science, University College TATI, Malaysia
ziti@uctati.edu.my https://orcid.org/0000-0001-6680-0009
[*]Corresponding Author

Article Info:

Article history:
Received date: 20.01.2026
Revised date: 19.02.2026
Accepted date: 19.03.2026
Published date: 30.03.2026

To cite this document:

Ali, M. A., Sulaiman, N. S., & Mohd Apandi, Z. F. (2026). A Review of Cyberbullying Detection on Social Media. *International Journal of Law, Government and Communication, 11*(43), 424-440.

DOI: 10.35631/IJLGC.1143028

Abstract:

The digital world is evolving quickly today. The way people communicate no longer relies on traditional methods like telegraphs, letters, and phone calls. The internet has transformed the global scene and affected social interaction, communication, and information sharing. This progress has led to the growth of digital social media platforms that enable all forms of knowledge sharing and communication, helping users. However, among these features, some users engage in risky behaviours such as cyberbullying and harassment. This paper explores cyberbullying, social media, and machine learning, outlining their definitions, categories, and functions. It reviews detection methods for cyberbullying, focusing on data sources, feature extraction, evaluation metrics, and classification techniques, especially Machine Learning and Natural Language Processing (NLP). The paper identifies gaps, assesses current approaches, and proposes future research directions.

Keywords:

Cyberbullying, Cyberbullying Detection, Machine Learning, NLP, Evaluation Metric

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

Introduction

One of the main ways teens communicate and gather information is through the Internet. In 2012, 81% of teens with computers used the Internet daily to interact with their classmates (Anderson et al., 2017). This rise in online activity prompted the development of social media (SM) applications that make it easy to share opinions, products, lifestyles, and more. SM includes various online platforms such as blogs, social networks, product reviews, video sharing, and virtual worlds (Asio & Khorasani, 2015). These platforms enable users to express and display opinions and emotional data through public or private online profiles.

Bullying is deliberate and hurtful behavior by an individual or group, often involving repeated physical, social, or verbal abuse, damage, or intimidation directed at a target (Pishchenko & Solovey, 2022). It includes various forms: physical bullying, such as hitting or pushing; verbal bullying, like name-calling or insults; social bullying, such as exclusion or spreading rumors; and cyberbullying, which uses digital platforms to harass or shame someone (Nazir & Thabassum, 2021). Cyberbullying has become a significant issue in the digital age, impacting people of all ages, especially teenagers. It involves using information and communication technology to harass, threaten, or humiliate others (Talpur & O'Sullivan, 2020).

This phenomenon can lead to severe psychological and emotional trauma. A survey revealed that 16% of high school students reported facing electronic bullying within a year, emphasizing its widespread occurrence (Samsudin et al., 2023). The anonymity and broad reach of the Internet make it easier for bullies to target individuals (Nunavath, V., 2024), often resulting in relentless and pervasive harassment.

This review addresses issues related to cyberbullying and examined achievements in extracting built-in features by comparing existing ones. It also provides an overview of the main challenges, current solutions, and future research directions. The remainder of the paper is organized as follows: Section Two outlines the review process. Conversely, Section Three offers an overview of feature extraction in cyberbullying detection algorithms, along with their advantages and disadvantages. Section Four discusses the difficulties and challenges involved in feature extraction, and the final section suggests directions for future research.

Literature Review

This section describes the literature review process, including searching, setting inclusion and exclusion criteria, and presenting the results. Several online archives, including IEEE Explore, Springer, ScienceDirect, ACM, ResearchGate, and Google Scholar, were used to search for specific terms and citations. Keywords such as cyberbullying, cyberbullying detection, Machine Learning, feature extraction, and evaluation metrics were used in searches to locate relevant papers. Only articles, journals, and doctoral theses published between 2019 and 2025 are included, as earlier publications are covered in previous surveys. Most articles published before 2019 are also included, as they provide a clear, foundational perspective. Abstracts,

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440
editorials, and unpublished content such as reports and theses are excluded. Publications written in languages other than English or lacking an English translation are also not included. In the first step, titles and abstracts are retrieved from the mentioned electronic sources through keyword searches to assess their relevance to the topic. The selected abstracts are reviewed, and a list of pertinent publications is created. Every paper likely to meet the criteria is downloaded, and its methodology, results, and conclusion sections are thoroughly examined. The main findings of each publication are summarized in Excel worksheets and combined to facilitate visual analysis. Each cited work is carefully critiqued, aiding future researchers in identifying directions for subsequent studies.

Social Media

Social media includes websites and online platforms that allow users to share information, create content, interact with others, and participate in virtual communities (Aichner et al., 2021). "Social Media" (SM) refers to online platforms designed to enable communication and social interaction among users. This software or platform facilitates communication between users and allows them to share information and opinions in various formats, such as text, images, links, and videos. As a result, it has become a vital communication environment in today's world and plays an important role in spreading information.

In another sense, social media can also be viewed as a digital marketing channel that allows advertisers to engage consumers (Appel et al., 2020). However, social media can be broadly seen as a digital space where people conduct many aspects of their lives, rather than just a digital medium and a set of specific technical services. People worldwide use social media in many ways, from reading the news to doing business and staying connected with friends and family.

Machine Learning

Machine Learning (ML) is the ability of systems to learn from training data specific to a problem and to automate the creation of analytical models to solve related tasks. (Janiesch et al., 2021) This ML can be enhanced through testing on the studied problems with combinations of solvers. By evaluating tests and results (experience) against a specific performance metric, machine learning algorithms can uncover hidden patterns and methods in incoming data without being explicitly programmed for the task.

Machine learning is a subfield of AI that enables systems to learn from data and improve over time. (Tong et al., 2025) It is broadly categorized into four types (Ray, 2019): Supervised learning, where systems use labeled data for prediction, such as image classification; Unsupervised learning, where systems identify patterns in unlabeled data, like clustering; and Semi-Supervised learning, which combines both labeled and unlabeled data for scenarios with limited labeling (Dash et al., 2021). While reinforcement learning focuses on optimizing actions through interaction with an environment and feedback, deep learning is more advanced and can automatically identify labels from data. It is primarily used in natural language processing and computer vision applications.

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

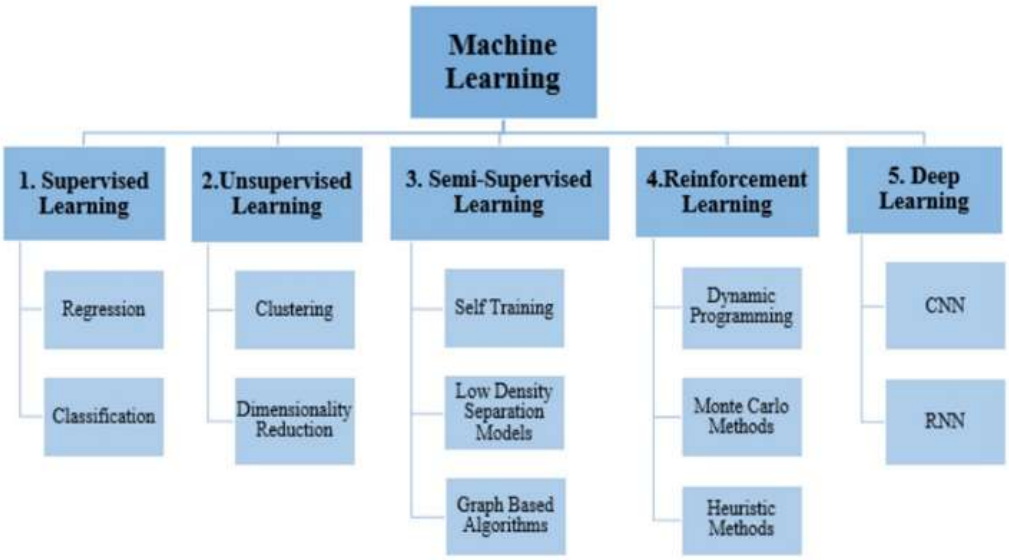


Figure 1: Types of Machine Learning Algorithms.
Source: Nassif, Ali Et Al. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review.

Machine learning algorithms are usually categorized based on their intended outcome or task. In Supervised Learning, the algorithm learns a function that maps inputs to desired outputs, often used in classification problems where the goal is to assign input data to one of several predefined categories based on labeled examples (Kadhim, 2019). Unsupervised Learning focuses on modeling a set of inputs without labeled examples, aiming to discover patterns or structures within the data. Semi-supervised learning combines labeled and unlabeled data to build a classifier or function, leveraging the strengths of both. Reinforcement Learning involves developing a policy to decide actions based on observations of the environment, where each action affects the environment, and feedback guides the learning process (Oladipupo Ayodele, 2010). Finally, Deep Learning enables the algorithm to learn its own inductive biases from prior experience, often using complex neural networks to identify patterns in large datasets (Nassif et al., 2019).

Survey Data Collection, Process, and Method

Survey data collection is essential in cyberbullying detection research because it gathers information about users' experiences, behaviors, and perceptions related to online harassment. Usually, data is collected using specific software, such as an API or a crawler, and with existing data obtained in raw form and cleaned for analysis of social media users. Researchers then perform statistical or descriptive analysis to identify patterns and insights that can support the development and improvement of cyberbullying detection models.

Data Collection

A researcher can choose different options for data collection from social media, each with its own advantages and challenges. There are two methods: primary, where data is collected directly from the source and has not been used before, and secondary, where data is gathered from existing sources and has been used in previous studies. Platforms like Twitter, Instagram, and Facebook that provide API tools enable structured, compliant retrieval of posts, comments, and user engagement metrics. However, most APIs are rate-limited and/or require

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

authentication, making them suitable for small-sample research or targeted data queries (Lomborg & Bechmann, 2014).

Web scraping serves as an alternative to data mining when API access is not available, directly extracting information from web pages. It works by sending HTTP requests to a website, parsing the HTML response, extracting relevant data, and storing it in a structured format. With a vast amount of online data—from social media and websites to portals and other platforms—efficiently gathering this information is essential for organizations to gain valuable insights. Web scraping is one of the primary methods used for this purpose (Lotfi et al., 2021).

Other ways to access social media data besides direct collection include third-party data providers and public datasets; however, these are often expensive and sometimes not customizable. More affordable options include academic repositories and secondary data sources, such as Google Dataset, Kaggle, and Mendeley Data, which offer open datasets suitable for historical analysis but may not be up to date. Each of these methods requires researchers to weigh the study's needs against ethical guidelines and available resources, balancing convenience, legality, and data completeness.

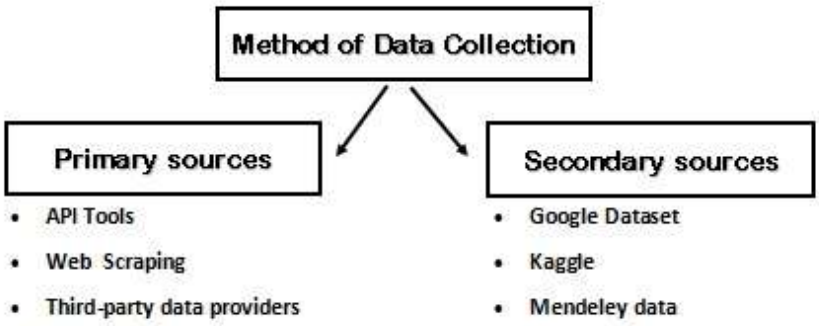


Figure 2: Method of Data Collection for Cyberbullying Detection Research.

Table 1: The Study Shows Differences in Data Collection Methods for Each Research.

Study	Year	Data Collection Type	Social Media Platform
Naf'an, M. Z., et al.	2019	Instagram API	Instagram
Theng, C. P. et al.	2021	Rapid Miner Studio	Twitter
Wang, Anqi & Potika, Katerina	2021	Twitter API	Twitter
Arnisha Akhter et al.	2019	Kaggle.com	Twitter
Perera, Andrea & Fernando, Pumudu	2024	Web scraping / Internet Archive	Twitter

Cyberbullying Detection Process

Common techniques for detecting cyberbullying include using Natural Language Processing (NLP) and Machine Learning methods (Van Hee et al, 2018). NLP tools such as sentiment analysis, text classification, and entity recognition are used to identify harmful language,

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

negative sentiment, and targeted attacks in online content. Using these tools and extractions, the machine learning model's classification algorithm will categorize content as bullying or non-bullying based on the training dataset and the tests performed on it. More subtle cases of bullying can be detected by supervised, semi-supervised, and deep learning models, such as recurrent neural networks and transformers (Rahman, 2022). These combined methods help analyze and classify online interactions to identify harmful behaviors. Data preprocessing, feature extraction, method selection, and training, validation, and testing for detection accuracy are all vital components for effective cyberbullying detection.

Pre-Processing of Data

The first step is data processing, which involves converting unstructured social media data into a suitable format and structure for analysis. First, irrelevant data points, such as duplicates and noise, are removed through the data cleaning (rubbish out) process (Fathoniah & Rozikin, 2022). Relevance filtering is essential because most social media data consist of spam, ads, and other irrelevant content that could skew analyses (Chandrasekaran et al., 2022). Additionally, the diversity of languages and variations in language use, including slang or abbreviations, often requires specific preprocessing depending on the use case. Techniques such as lemmatization, stemming, and machine translation (which return words to their root forms) are vital for maintaining accuracy.

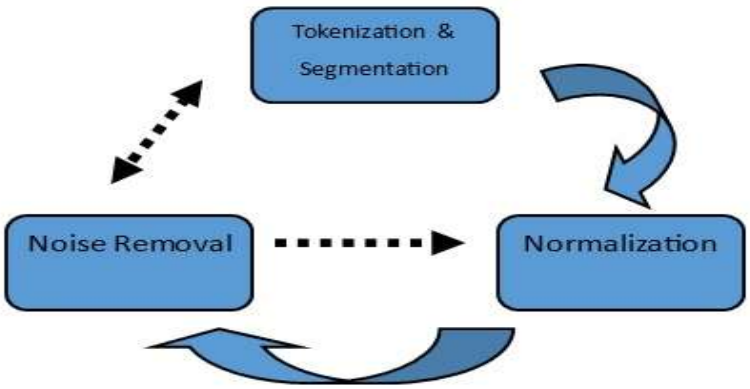


Figure 1: Represents The Most Common Framework Text Pre-Processing Data Process.

Text pre-processing in NLP includes steps like tokenization, stop-word removal, and lemmatization. These steps prepare the text for analysis by breaking it into components and normalizing it. This process is essential for sentiment analysis, topic modeling, and text-based analytics. analytics.

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

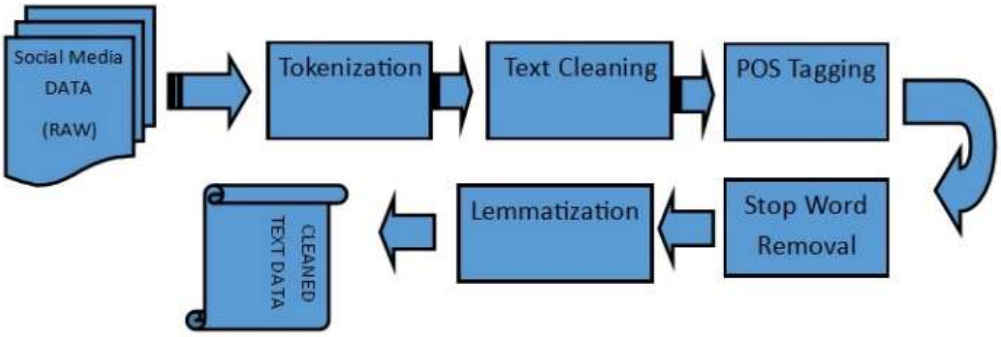


Figure 1: The Text Pre-Processing Step in Natural Language Processing.
Source: Chai Cp. (2023). Comparison of Text Pre-processing Methods in Natural Language Engineering.

Feature Extraction

An important step in cyberbullying detection is feature extraction, which transforms unstructured text into meaningful representations for machine learning models. To identify bullying content, a variety of features are analyzed, including lexical, syntactic, semantic, psycholinguistic, user behavior, and stylistic features (Sultan et al., 2023).

Lexical features include word-based patterns like counting word frequency, slang use, and n-grams, while syntactic features analyze sentence structure using Part-of-Speech (POS) tagging and dependency parsing. Semantic features such as TF-IDF, word embeddings (Word2Vec, GloVe, BERT), and sentiment analysis help understand the context and meaning of words. Psycholinguistic tools such as LIWC and emotion lexicons can further detect aggression and negative emotions in messages (Teng et al., 2023).

User behavior and network features track user interactions and message frequency, while stylistic features like excessive capitalization, punctuation, and character repetition can indicate aggressive intent. Machine learning models can effectively identify cyberbullying in online communications by extracting and combining these features.

Algorithms and Model

Cyberbullying detection depends heavily on choosing the right machine learning model to classify abusive content. Generally, there are two types: traditional machine learning and deep learning models (Muneer & Fati, 2020). This model, currently used for cyberbullying detection, mainly relies on available data for analysis.

Traditional models such as Random Forests, Support Vector Machines, Naive Bayes, and Logistic Regression work well for small datasets or simpler tasks. These models rely on manually engineered features, such as word counts or TF-IDF values (Sultan et al., 2023). They are faster, more interpretable, and require less computational power. However, they may struggle with complex or subtle text.

While deep learning models like CNNs, RNNs, and LSTMs are much more effective with large-scale language datasets, they handle complex patterns significantly better (Sultan et al., 2023). These models capture most of the text's contextual and sequential relationships and are

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

mainly skilled at attending to the finer details of language. They require additional computational resources, large datasets, and substantial training data.

Performance Metrics

When assessing the performance of cyberbullying detection models, various metrics are commonly used to evaluate their effectiveness. These metrics measure how accurately a model identifies harmful content while minimizing errors and balancing multiple predictive factors.
Accuracy is a key performance indicator that measures how correct a model is by calculating the ratio of correct predictions—true positives and false negatives—out of all predictions. While accuracy offers a broad view of the model's training success, it might not be sufficient as a performance measure, especially with imbalanced datasets (Željko Đ. Vujovic, 2021). See the formula:

$$Accuracy: x = \frac{\text{true}(+) + true\ (-)}{\text{true}(\mp) + true(-) + false(+) + false(-)}$$

In such cases, high accuracy can mislead a model, where instances of cyberbullying might be very few compared to non-bullying content. The algorithm could achieve high accuracy mainly by predicting the majority class (nonbullying), even if it has not learned to identify bullying content effectively.
Precision measures how accurately a model detects cyberbullying cases, especially the number of true positives, which is the ratio of correctly identified cyberbullying instances to all predicted cases (Wan Ali et al., 2018). Precision is especially important when the cost of errors is high, as false positives can mistakenly be classified as bullying, leading to negative consequences.

$$Precision: P = \frac{\text{true}(+)}{\text{true}(\mp) + false(+)}$$

The higher the precision, the more it shows that a model is mostly accurate in labeling something as cyberbullying.
It represents sensitivity, also called the true positive rate, which is the proportion of correctly identified cyberbullying cases. It shows the ratio of true positives to all bullying cases. Recall becomes important when false negatives are unacceptable, harmful content is missed, and additional bullying could occur (Rainio et al., 2024; Vujović, 2021).

$$Recall: R = \frac{\text{true}(+)}{\text{true}(\mp) + false(-)}$$

A high reminiscing score indicates that the model detects most bullying cases, reducing the risk of missing harmful content.
The F1-score is the harmonic mean of precision and recall, effectively balancing false positives and false negatives. It is particularly useful in cases of data imbalance, where one class is much more common than another, such as cyberbullying compared to non-bullying content.

$$F - score: Fx = (1 + x^2)\frac{Precision * Recall}{x^2 * Precision + Recall}$$

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

The F1-score effectively balances precision and recall, assessing a model's performance across multiple metrics and preventing false positives or false negatives from dominating the outcomes (Rainio et al., 2024). This is important in situations where both errors matter equally, such as in cyberbullying.

Discussion

Sentiment analysis is a technique for classifying opinions expressed on social media. It aims to determine whether sentences are positive, negative, or neutral, which is very helpful for identifying aggressive content that may constitute cyberbullying (Naf'an, M. Z., et al., 2019; Theng et al., 2021). According to Hamiza Wan Ali et al. (2018), Natural Language Processing (NLP) and machine learning are well-known methods for detecting bullying-related keywords in a dataset, while Support Vector Machine (SVM) is effective for classifying highly imbalanced text, such as identifying cyberbullying based on content features.

Theng et al. (2021) noted that the Support Vector Machine model is more reliable than Naive Bayes and K-Nearest Neighbors (k-NN). This may be because the K-Nearest Neighbors (k-NN) algorithm relies on statistics and comparisons, which require analyzing large feature sets. The Support Vector Machine performs offline learning to determine the optimal hyperplane. It relies on the training set to find an equation that separates the two classes, or a hyperplane. Although SVM and Naive Bayes are effective for high-dimensional data analysis (Sultan et al., 2023), Logistic Regression is quite suitable for linear problems, and Random Forests can effectively handle diverse feature types.

The LSTM model has been proposed to address the vanishing gradient problem often faced by traditional RNNs, allowing it to remember information over longer sequences. It uses dynamic gates, or memory cells, to regulate the flow of information, helping the model determine which parts of the input are relevant and should be remembered or forgotten during training. In the (Bokolo & Liu, 2023) paper, cyberbullying detection leverages the effectiveness of Bi-LSTM in capturing the complexity of human languages.

In contrast, (Rahman, 2022; Sultan et al., 2023) and (Engel, E., 2023) recently developed cyberbullying detection modules using deep learning models, such as transformers (e.g., BERT), which achieve state-of-the-art results but require resource-intensive datasets. Conversely, Aliyeva et al. (2024) used a Multi-Layer Perceptron (MLP) to solve the well-known XOR problem, a significant hurdle in early neural network research. It has proven highly effective in classification and generalization tasks. After pre-processing and feature engineering, noise is removed from the data, and the most relevant features are fed into the MLP's dense layers, enabling optimal performance.

In summary, three categories of Machine Learning models—Supervised, Semi-Supervised, and Deep Learning—are commonly used to detect cyberbullying with existing datasets and technologies. Most of the modules rely heavily on the data source and the computer's capabilities and processing speed for detection. This work emphasizes the importance of evaluating these algorithms against performance metrics such as accuracy, precision, recall, and F1-score to ensure their effectiveness in identifying cyberbullying on social media platforms.

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

Table 1: Summary of Study on Cyberbullying Detection

Authors	Year	Dataset	ML type	Classifier and Feature Extraction Model	Result Matric
Naf'an, M. Z., et al.	2019	Instagram	Supervised Learning	Naïve Bayes Classifier TF-IDF method (Term Frequency - Inverse Document Frequency) K-Fold Cross-Validation Method	K-Fold Cross Validation Accuracy matric 83.53%
Theng, C. P. et al.	2021	Twitter	Supervised Learning	Naïve Bayes Classifier TF-IDF method (Term Frequency - Inverse Document Frequency) K-Fold Cross-Validation Method	List of different words SVM 60% KNN 52.84% Naïve Bayes 50.16%
Wang, Anqi & Potika, Katerina	2021	Twitter	Supervised Learning	Random Forest SVM (linear) Logistic Regression Adaptive Boosting (AdaBoost) Naïve Bayes CNN	Support Vector Machine achieved 89% accuracy
Arnisha Akhter et al.	2019	Twitter	Supervised Learning	Multinomial Naïve Bayes + Fuzzy Logic	Accuracy rate of 88.89% vs SVM 76.38%
Perera, Andrea & Fernando, Pumudu	2024	Twitter	Supervised Learning	SVM, Naïve Bayes Logistic Regression	Combined features (Tf-idf, Sentiment core) Acuray 75.17%, Precision 75%, Recall 75%, F1-measure 75%, & LR Accuracy 78.52%
Wang, Jason et al.	2020	Twitter	Semi-Supervised Learning	Dynamic Query Expansion (DQE) Graph Convolutional Network (GNN)	BOW+XGBoost Accuracy 94.38% TF-IDF+XGBoost F1-measure 94.44%

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

Aditya Desai et al.	2021	Twitter	Semi-Supervised Learning	Confusion matrix labels indicate Bullying (0) and non-Bullying (1). BERT	Accuracy of BERT Model 91.90%
B. G. Bokolo and Q. Liu	2023	Twitter	Semi-Supervised Learning + Deep Learning	Bidirectional Long Short-Term Memory (Bi-LSTM) Recurrent Neural Network (RNN)	Bi-LSTM 98% accuracy, SVM 97% accuracy, and Naive Bayes 85%
Alabdulwahab, Aljwharah et al.	2023	Twitter	Deep Learning	CNN-LSTM	Accuracy KNN 90%, SVM 92%, CNN-LSTM 96%
Aliyeva, Ç.O., Yağanoğlu, M	2024	Twitter	Deep Learning	MLP + TF-IDF + Unigram	TF-IDF and unigram methods Chi-Square method Accuracy 93.2%

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

Future Directions

Detecting cyberbullying is a crucial area of research, especially with the rising rate of online harassment and its serious effects. An effective detection system relies on feature extraction to identify relevant data traits needed to train a machine learning model. However, this process faces several challenges. One significant challenge is understanding context, as cyberbullying often involves indirect communication, sarcasm, or subtle signals that require advanced NLP techniques. Furthermore, the multimodal nature of online content-text, images, audio, and video-complicates feature extraction and demands sophisticated methods to integrate different data types (Rosa et al., 2019).

Online communication evolves rapidly, with slang, emojis, and trends constantly changing, making it more complex. Another issue is content ambiguity, which depends heavily on interpretation (Emmery et al., 2021). There is also a data imbalance problem: instances of bullying are rare compared to harmless content, making it difficult to detect minority patterns. Additionally, the noisy nature of online data-often filled with misspellings, grammatical errors, and irrelevant info-necessitates thorough preprocessing to extract useful features (Hasan et al., 2023). Privacy concerns and ethical restrictions further limit access to personal or sensitive data, which constrains feature extraction. The variation in expressions across regions and languages adds yet another layer of difficulty, requiring customized approaches.

Overcoming these challenges is crucial for creating accurate and trustworthy cyberbullying detection systems. Future research should aim to improve model accuracy by using larger, more diverse datasets from multiple social media platforms. Researchers can investigate advanced feature-extraction methods and contextual language analysis to understand online communication patterns better and enhance the detection of harmful content. Additionally, integrating deep learning and multimodal approaches that combine text, images, and emojis could increase system effectiveness. Other valuable directions include testing models across different platforms, applying explainable AI techniques to boost transparency, deploying real-time detection systems, and incorporating user behaviors and network metadata to improve prediction accuracy and promote safer online spaces.

Conclusion

The rapid development of technology has facilitated the growth of human communication networks, particularly through social media platforms. Conversely, when people misuse social media to bully others anonymously, they can be seen as uncivilized citizens. Generally, researchers have focused on identifying bullying keywords within texts by using text classification in Natural Language Processing (NLP) and machine learning techniques. The text describes the evolution of machine learning models for cyberbullying detection, highlighting the strengths of specific algorithms and stressing the importance of model evaluation to ensure effective performance.

Several machines and deep learning models have improved the detection of cyberbullying on social media platforms. Techniques such as sentiment analysis, natural language processing, and machine learning algorithms—including support vector machines, logistic regression, and random forests—offer reliable methods for identifying offensive content. Among these models, SVM is advantageous because it handles high-dimensional data well and outperforms other methods such as Naive Bayes and K-Nearest Neighbors. Recently, deep learning models such

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

as LSTMs and transformer-based models like BERT have excelled at recognizing complex language patterns. However, most of these models require large datasets, which need significant computational resources.

Nonetheless, while these newer methods offer significant promise, they will always need to be evaluated using performance metrics such as accuracy, precision, recall, and F1-score, which remain crucial to the reliability of our assessment. In the future, effective combinations of supervised, semi-supervised, and deep learning methods will need to be strategically considered for further efforts against online cyberbullying.

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

Acknowledgements: The authors sincerely thank all individuals and institutions who contributed to the completion of this study. We also appreciate our supervisors and academic mentors for their ongoing guidance, valuable suggestions, and constructive feedback throughout the research process. We also thank our institution for providing the academic resources and research environment that allowed us to complete this review. Finally, we would like to thank the researchers and scholars whose previous studies on cyberbullying detection and social media analysis offered the foundation and inspiration for this work.

Funding Statement: No Funding

Conflict of Interest Statement: The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have contributed to this work and approved the final version of the manuscript for submission to the International Journal of Law, Government and Communication (IJLGC).

Ethics Statement: This study did not involve any human participants, animals, or sensitive data that required ethical approval. The authors confirm that the research was conducted in accordance with accepted standards of academic integrity and ethical publishing.

Author Contribution Statement: All authors played a vital role in developing this manuscript. M. Azizuddin was responsible for the conceptualization, methodology, and overall supervision of the study, along with data collection, analysis, and interpretation of results. N. Suhana and Ziti F. contributed to the literature review, drafting, and critical revision of the manuscript. All authors reviewed and approved the final version before submission.

International Journal of Law, Government and Communication
EISSN : 0128-1763
Volume: 11 Issue: 43 (March 2026) PP. 424-440

References

Aditya Desai, Shashank Kalaskar, Omkar Kumbhar, Rashmi Dhumal (2021). Cyber Bullying Detection on Social Media Using Machine Learning. *ITM Web Conf. 40 03038* https://doi.org/10.1051/itmconf/20214003038

Alabdulwahab, Aljwharah & Haq, Mohd Anul & Alshehri, Mohammed. (2023). Cyberbullying Detection using Machine Learning and Deep Learning. *International Journal of Advanced Computer Science and Applications. 14. 424–432*. https://doi.org/10.14569/IJACSA.2023.0141045

Aliyeva, Ç.O., Yağanoğlu, M. (2024). Deep learning approach to detect cyberbullying on Twitter. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-19869-3

Appel, G., Grewal, L., Hadi, R. et al. (2020). The future of social media in marketing. *J. of the Acad. Mark. Sci. 48, 79–95*. https://doi.org/10.1007/s11747-019-00695-1

Arnisha Akhter, Uzzal K. Acharjee, Md Masbaul A. Polash. (2019). Cyber Bullying Detection and Classification using Multinomial Naïve Bayes and Fuzzy Logic. *I.J. Mathematical Sciences and Computing, 2019, 4, 1-12.* https://dx.doi.org/10.5815/ijmsc.2019.04.01

Asio, S. M., & Khorasani, S. T. (2015). Social media: A platform for innovation. *IIE Annual Conference and Expo 2015*.

B. G. Bokolo and Q. Liu (2023). Cyberbullying Detection on Social Media Using Machine Learning. *IEEE INFOCOM 2023 - Conference on Computer Communications Workshops, INFOCOM WKSHPS 2023,* https://dx.doi.org/10.1109/INFOCOMWKSHPS57453.2023.10226114

Chai CP. (2023). Comparison of text preprocessing methods. *Natural Language Engineering. 2023;29(3):509-553*. https://doi:10.1017/S1351324922000213

Dash, Sushree & Nayak, Subrat & Mishra, Debahuti. (2021). A Review on Machine Learning Algorithms. *Smart Innovation, Systems and Technologies (2021).* https://doi.org/10.1007/978-981-15-6202-0_51

Emma Louise Anderson, Eloisa Steen & Vasileios Stavropoulos. (2017). Internet Use and Problematic Internet Use: a systematic review of longitudinal research trends in adolescence and emergent adulthood, *International Journal of Adolescence and Youth, 22:4, 430-454*. https://doi.org/10.1080/02673843.2016.1227716

Emmery, C., Verhoeven, B., De Pauw, G. et al. (2021). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. *Lang Resources & Evaluation 55, 597–633*. https://doi.org/10.1007/s10579-020-09509-1

Fathoniah, S., & Rozikin, C. (2022). Analisis Sentimen Masyarakat terhadap Teroris dalam Media Sosial Twitter menggunakan NLP. *Jurnal Ilmiah Wahana Pendidikan, 8(13).*

G. Chandrasekaran, N. Antoanela, G. Andrei, C. Monica and J. Hemanth, (2022). Visual Sentiment Analysis Using Deep Learning Models with Social Media Data. *Applied Sciences, 12(3).* https://dx.doi.org/10.3390/app12031030

H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J.P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A.M. Veiga Simão, I. Trancoso (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior, Volume 93,2019, Pages 333–345*, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2018.12.021

Hasan, M. T., Hossain, M. A. E., Mukta, M. S. H., Akter, A., Ahmed, M., & Islam, S. (2023). A Review on Deep-Learning-Based Cyberbullying Detection. *Future Internet, 15(5), 179*. https://doi.org/10.3390/fi15050179

Janiesch, C., Zschech, P. & Heinrich, K. (2021). Machine learning and deep learning. *Electron Markets 31, 685–695*. https://doi.org/10.1007/s12525-021-00475-2

Kadhim, A.I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artif Intell Rev 52, 273–292*. https://doi.org/10.1007/s10462-018-09677-1

Lomborg, S., & Bechmann, A. (2017). Using APIs for Data Collection on Social Media. *Information Society, 30(4)*. https://doi.org/10.1080/01972243.2014.915276

Lotfi, Chaimaa & Srinivasan, Swetha & Ertz, Myriam & Latrous, Imen. (2021). Web Scraping Techniques and Applications: A Literature Review. *SCRS Conference Proceedings On Intelligent Systems.* https://doi.org/10.52458/978-93-91842-08-6-38

Md. Habeeb Ur Rahman, Mudigonda Divya, B. Ramya Reddy, Dr. K Sateesh Kumar, P. Ramya Vani. (2022). Cyberbullying Detection using Natural Language Processing. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 10(V), 5241–5248.* https://doi.org/10.22214/ijraset.2022.43683

Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on Twitter. *Future Internet, 12(11).* https://doi.org/10.3390/fi12110187

Naf'an, M. Z., Bimantara, A. A., Larasati, A., Risondang, E. M., & Nugraha, N. A. S. (2019). Sentiment Analysis of Cyberbullying on Instagram User Comments. *Journal of Data Science and Its Applications, 2(1), 38-48*. https://dx.doi.org/10.21108/jdsa.2019.2.20

Nassif, Ali & Shahin, Ismail & Attili, Imtinan & Azzeh, Mohammad & Shaalan, Khaled. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access. PP. 1–1*. https://doi.org/10.1109/ACCESS.2019.2896880

Nazir, Thseen & Thabassum, Liyana. (2021). Cyberbullying: Definition, Types, Effects, Related Factors and Precautions to Be Taken During COVID-19 Pandemic. *The International Journal of Indian Psychology. 09. 480-491*. https://doi.org/10.25215/0904.047

Nunavath, V. (2024). A Hybrid Deep Learning Approach for Multi-Class Cyberbullying Classification Using Multi-Modal Social Media Data. *Applied Sciences, 14(24),* 12007. https://dx.doi.org/10.3390/app142412007

Oladipupo, T. (2010). Types of Machine Learning Algorithms. *InTech*. *New Advances in Machine Learning*. https://doi.org/10.5772/9385

Perera, Andrea & Fernando, Pumudu. (2024). Cyberbullying Detection System on Social Media Using Supervised Machine Learning. *Procedia Computer Science. 239. 506-516.* https://doi.org/10.1016/j.procs.2024.06.200

Pishchenko, Hennadiy & Solovey, Olena. (2022). Bullying As An Object Of Criminological Research. *Criminalistics and Forensics*. https://doi.org/10.33994/kndise.2022.67.20

PMID: 30296299; PMCID: PMC6175271.

Rainio, O., Teuho, J. & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Sci Rep 14, 6086*. https://doi.org/10.1038/s41598-024-56706-x

Ray, Susmita. (2019). A Quick Review of Machine Learning Algorithms. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019 35-39*. https://doi.org/10.1109/COMITCon.2019.8862451

Samsudin EZ, Yaacob SS, Xin Wee C, Mat Ruzlin AN, Azzani M, Jamil AT, Muzaini K, Ibrahim K, Suddin LS, Selamat MI, Ahmad Saman MS, Abdullah NN, Ismail N, Yasin SM, Azhar ZI, Ismail Z, Rodi Isa M, Mohamad M (2023). Prevalence of cyberbullying victimisation and its association with family dysfunction, health behaviour and psychological distress among young adults in urban Selangor, Malaysia: a cross-sectional study. *BMJ Open. 2023 Nov 15;13(11):e072801*. https://doi.org/10.1136/bmjopen-2023-072801

Sultan, D., Omarov, B., Kozhamkulova, Z., Kazbekova, G., Alimzhanova, L., Dautbayeva, A., Zholdassov, Y., & Abdrakhmanov, R. (2023). A Review of Machine Learning Techniques in Cyberbullying Detection. *In Computers, Materials and Continua (Vol. 74, Issue 3).* https://doi.org/10.32604/cmc.2023.033682

T. H. Teng & K. D. Varathan (2023). Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches. *IEEE Access, vol. 11, pp. 55533-55560, 2023*, https://doi.org/10.1109/ACCESS.2023.3275130

Talpur BA, O'Sullivan D (2020). Cyberbullying severity detection: A machine learning approach. *PLOS ONE 15(10): e0240924*. https://doi.org/10.1371/journal.pone.0240924

Theng, C. P., Othman, N. F., Syahirah, R., Anawar, S., Ayop, Z., & Ramli, S. N. (2021). Cyberbullying detection in Twitter using sentiment analysis. *IJCSNS International Journal of Computer Science and Network Security, VOL.21 No.11* . https://doi.org/10.22937/IJCSNS.2021.21.11.1

Thomas Aichner, Matthias Grünfelder, Oswin Maurer, and Deni Jegeni (2021).Twenty-Five Years of Social Media: A Review of Social Media Applications and Definitions from 1994 to 2019.*Cyberpsychology, Behavior, and Social Networking.(Apr 2021).215-222.* http://doi.org/10.1089/cyber.2020.0134

Tong, M., Zhou, J., Akkaya, Z., Majumdar, S., & Bhattacharjee, R. (2025). Artificial intelligence in musculoskeletal applications: *A primer for radiologists*. https://doi.org/10.4274/dir.2024.242830

Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, Verhoeven B, De Pauw G, Daelemans W, Hoste V(2018). Automatic detection of cyberbullying in a social media text. *PLoS One. 2018 Oct 8;13(10): e0203794*. https://doi.org/10.1371/journal.pone.0203794

Wan Ali, Wan Noor Hamiza & Mohd, Masnizah & Fauzi, Fariza. (2018). Cyberbullying Detection: An Overview. 1-3. 10.1109/CR.2018.8626869). *Proceedings of the 2018 Cyber Resilience Conference, CRC 2018.* https://dx.doi.org/10.1109/CR.2018.8626869

Wang, Anqi & Potika, Katerina. (2021). Cyberbullying Classification based on Social Network Analysis. *Proceedings - IEEE 7th International Conference on Big Data Computing Service and Applications, BigDataService 2021*. https://doi.org/87-95. 10.1109/BigDataService52369.2021.00016

Wang, Jason & Fu, Kaiqun & Lu, Chang-Tien. (2020). SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection. 1699-1708. https://doi.org/10.1109/BigData50022.2020.9378065

Željko Đ. Vujovic (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications (IJACSA), 12(6)*. http://dx.doi.org/10.14569/IJACSA.2021.0120670