UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Cyberbullying Detection on Social Media: A Review of Features and Machine Learning Approaches

Che Syabila Che Shamsul[1*] & Wan Noor Hamiza Wan Ali[2]
Faculty of Artificial Intelligence
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia
Email: chesyabila@graduate.utm.my[1]; wannoorhamiza@utm.my[2]

Submitted: 15/9/2025. Revised edition: 28/11/2025. Accepted: 5/1/2026. Published online: 28/7/2026
DOI: https://doi.org/10.11113/ijic.v16n1-2.687

***Abstract*—The rapid growth of digital communication platforms has facilitated global connectivity while simultaneously intensifying the prevalence of cyberbullying, a form of online aggression with severe psychological consequences for victims. The research problem addressed in this paper concerns the limitations of manual moderation methods in detecting harmful online behavior at scale. Consequently, this review paper provides a comprehensive synthesis of recent research on cyberbullying detection, with an emphasis on feature engineering and supervised learning algorithms, including Support Vector Machine (SVM) and Random Forest. Feature engineering strategies that include lexical, sentiment, user, and network-based features are examined for their role in enhancing detection accuracy. Analysis of 13 key studies reveals that SVM and Random Forest consistently outperform other classifiers, especially when combined with robust optimization techniques. The review identifies that content-based features, particularly TF-IDF, combined with hyperparameter tuning, achieve accuracies exceeding 94% in optimal configurations. The study concludes that while current models demonstrate substantial progress, challenges related to linguistic ambiguity, data imbalance, cross-platform generalizability, and lack of multilingual support remain critical directions for future research.**

***Keywords*—Cyberbullying Detection, Machine Learning, Supervised, Social Media Analysis, Hyperparameter Tuning**

I. INTRODUCTION

The widespread use of digital platforms has transformed how individuals interact, communicate, and share information with others. Social media platforms such as Twitter, Instagram, TikTok, and Facebook have enabled unprecedented connectivity, but this accessibility has also facilitated the rise of online harassment and cyberbullying [14]. Cyberbullying is defined as intentional, repetitive, and aggressive behavior conducted via digital means, often targeting vulnerable individuals who are unable to defend themselves [13], [15]. Unlike traditional bullying, its reach is not constrained by specific time or space, making its impact both persistent and far-reaching.

The consequences of cyberbullying can be severe, ranging from anxiety and depression to self-harm and suicide among adolescents and young adults. Studies reveal that between 14.6% and 52.2% of internet users have experienced some form of victimization, while perpetration rates range between 6.3% and 32% [16]. In Malaysia, prevalence rates are notably high, with recent studies reporting that 79.5% of students had been victims of cyberbullying [17].

Traditional moderation approaches, such as manual reporting or platform-based flagging, are inadequate because of the scale, speed, and linguistic complexity of online communication. Consequently, researchers have increasingly turned to machine learning techniques for automated cyberbullying detection [18], [19]. Among the many algorithms applied, Support Vector Machines (SVM) and Random Forest are widely used for their ability to handle high-dimensional text data and detect subtle patterns of abusive language.

This review paper explores the development of cyberbullying detection models in the literature, focusing on content-based features and supervised learning algorithms. It provides comparative evaluation studies, highlights methodological strengths and gaps, and positions a recent project utilizing several algorithms with optimized features as a practical contribution to this field.

To structure this review clearly, Section 2 outlines the methodology, detailing the selection criteria and review process applied to the chosen studies. Section 3 presents a detailed synthesis of current work on cyberbullying detection,

covering social media behavior, feature engineering, and algorithm approaches. Section 4 discusses limitations and research gaps, and Section 5 concludes with future research directions and implications for cyberbullying detection using machine learning.

II. METHODOLOGY

This section describes the approach used to conduct the literature review for this study. This study applied a Systematic Literature Review (SLR) framework to analyze machine learning on cyberbullying detection models published between 2020 and 2025. The search was conducted across multiple academic databases, including IEEE Xplore, Scopus, ScienceDirect, SpringerLink, and the ACM Digital Library. Additional databases were also explored to ensure comprehensive coverage of relevant interdisciplinary research. Keywords such as "cyberbullying detection", "machine learning", "SVM", "Random Forest", "feature engineering", and "social media" were used in combination with Boolean operators to refine the results. Studies were included if they explicitly addressed cyberbullying or online harassment detection, utilized social media text datasets, employed supervised machine learning algorithms, and provided clear methodology, feature extraction techniques, and performance metrics. Conversely, studies were excluded if they were published in languages other than English, conceptual discussions or surveys, and deep learning studies without comparison to traditional algorithms. The initial search identified 83 articles. After reviewing titles and abstracts, 47 studies satisfied the inclusion requirements. Following a thorough full-text evaluation, 13 studies were selected for detailed analysis based on their methodological strength and the comprehensiveness of their reported results.

III. CYBERBULLYING DETECTION

A. Cyberbullying on Social Media

Cyberbullying takes many forms, including offensive name-calling, spreading rumors, threats, stalking, and the sharing of explicit images. The online environment, with its anonymity and rapid virality, amplifies these harms beyond those typically observed in traditional bullying [20], [21]. Adolescents and minority groups are disproportionately impacted due to their extensive engagement on social media platforms and their vulnerability to peer influence [22].

Current detection approaches face significant challenges that limit their real-world effectiveness. A major shortcoming lies in their limited capacity to generalize across different social media environments, as most existing studies are centered primarily on Twitter datasets while ignoring distinct linguistic patterns and communication norms characteristic found on Instagram, TikTok, or Facebook [23]. Furthermore, the majority of existing research is confined to English-language datasets, reducing the models' effectiveness in multilingual and cross-cultural settings where expressions of cyberbullying may vary considerably [18].

Another critical challenge is that the linguistic complexity of online interactions complicates detection efforts. Sarcasm, coded language, emojis, and abbreviations like "lol" or "ASAP" often conceal harmful intent [37]. A phrase that appears harmless in one context may carry malicious intent in another [25]. This context-dependent nature of language underscores the limitations of traditional lexicon-based approaches and highlights the necessity for advanced natural language processing (NLP). Although certain models are effective at identifying overt profanity or explicit keywords, they often fail to recognize subtle or indirect aggression, such as sarcasm or platform-specific coded language, which are major obstacles to real-world [13], [26]. The challenge of detection is made more difficult by the inconsistent nature of abuse across different platforms. For example, Twitter often involves concise, aggressive statements, while Instagram is prone to body shaming and symbolic emojis in the comment section.

Additionally, the severe class imbalance present in cyberbullying datasets, where harmful content represents a small percentage of total posts, often results in models that are biased toward majority classes, achieving high overall accuracy while failing to detect minority instances of abuse effectively [13]. These constraints highlight the need for more robust, adaptive, and context-aware approaches that can operate reliably across diverse platforms, languages, and temporal contexts.

B. Features Used in Cyberbullying Detection

A comprehensive understanding of cyberbullying behavior is crucial for building effective detection systems. Feature engineering is especially important since the selection and extraction of meaningful features directly influence how well a model performs. Modern detection approaches generally rely on four major categories of features: content-based, sentiment-based, user-based, and network-based [27]. Each of these categories captures different dimensions of cyberbullying, enabling a more thorough analysis of harmful interactions within social media environments.

Content-based features are the foundation of most cyberbullying detection systems, as they directly analyze the textual content produced on social media platforms. Among the most widely used techniques is Term Frequency-Inverse Document Frequency (TF-IDF), which captures the relative importance of terms across documents and helps models distinguish between common words and discriminative vocabulary [4], [12]. Similarly, the Bag of Words (Bow) approach, though simple, has been effective in representing text by considering word occurrence frequencies without regard to order [1]. Another critical representation is the use of n-grams, which capture short sequences of words or characters and are particularly useful for modelling context and phrase structures [2].

Beyond these lexical features, researchers often include profanity indicators to identify the presence of abusive or explicit language [8] and pronoun frequency, especially second-person pronouns like "you" or "your", which frequently signal targeted aggression [5]. Studies consistently show that combining these surface-level lexical cues with deeper synthetic and semantic representations enhances overall model performance and allows for the detection of both explicit and implicit forms of harassment [28], [29].

While content-based features capture lexical and structural aspects of text, sentiment-based features provide an additional layer of insight by evaluating the emotional tone embedded in communication. Cyberbullying messages often exhibit strong negative polarity, heightened emotional intensity, and explicit expressions of anger, contempt, or hatred [30]. Such attributes differentiate bullying discourse from neutral or casual interactions. However, sentiment is not always straightforward to measure because sarcasm, irony, and backhanded compliments frequently mask harmful intent. This limitation has encouraged the development of contextual sentiment analysis methods that can capture implicit aggression and manipulative communication strategies.

In addition to text and sentiment, user-based features offer valuable contextual signals for cyberbullying detection. These features examine account-related information such as posting frequency, profile metadata, and behavioral patterns across time [31]. For instance, a sudden spike in activity during late hours might indicate malicious intent, while unusual posting behavior may serve as an early warning ofabusive conduct. Other user attributes, such as account age, verification status, and history of prior moderation, can also contribute to the assessment of credibility and likelihood of harmful behavior [23].

Finally, network-based features on structural relationships and interactions within social platforms are used to identify aggressor-victim dynamics. By analyzing how users interact within social graphs, such as the density of connections, clustering behavior, and centrality of user accounts, researchers can uncover hidden structures that are strongly correlated with bullying incidents [24]. For example, coordinated attacks may arise in tightly connected communities, while other cases may appear as isolated posts within broader interactional and relational contexts. Network-based features, therefore, serve as an important complement to other feature categories, enabling cyberbullying detection systems to move beyond individual text analysis and capture the collective dynamics of online harassment.

C. *Machine Learning Algorithms*

Machine learning techniques have been extensively explored and implemented for detecting cyberbullying within social media text [32]. Among these, supervised learning approaches have demonstrated particularly strong performance, as they rely on labeled datasets to recognize patterns linked to harmful content [2]. By analyzing textual data, such as tweets or online comments, these models can distinguish between bullying and non-bullying discourse, enabling automated classification, pattern recognition, and accurate predictions on previously unseen inputs. Since cyberbullying frequently appears in the form of text-based interactions, machine learning provides an effective means of processing the vast amounts of unstructured data generated on platforms like Twitter and Facebook. Within this context, algorithms such as SVM, Random Forest, Naïve Bayes, Logistic Regression, and Decision Tree are commonly employed to build robust detection systems. Fig. 1 displays the frequency of machine learning algorithms used by 13 selected studies on cyberbullying detection.

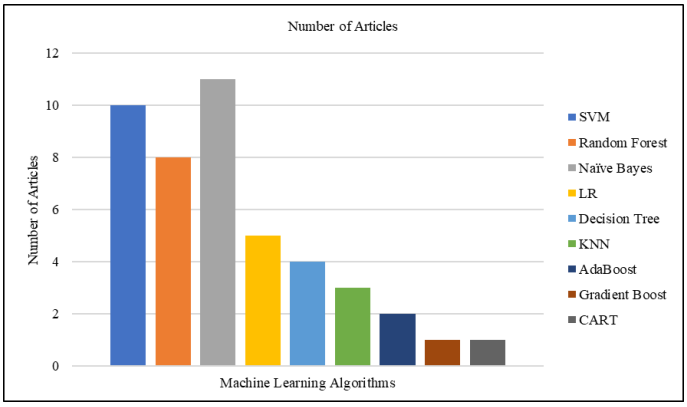

Fig. 1. The usage frequency of various machine learning algorithms in selected articles

SVM is one of the most widely applied algorithms in cyberbullying detection because of its ability to manage high-dimensional feature spaces efficiently. Its strength lies in creating optimal decision boundaries, which makes it particularly effective in distinguishing between subtle variations of online communication. Studies have demonstrated that SVM performs well in identifying nuanced and context-dependent language, making it a strong candidate for detecting forms of bullying that rely on implicit aggression rather than overt profanity [8], [9].

Random Forest (RF) is another powerful classifier frequently employed in cyberbullying research. As an ensemble of decision trees, Random Forest aggregates multiple weak learners to produce more stable and accurate predictions. It is especially valued for its resistance to overfitting and its ability to handle noisy and imbalanced datasets, which are common challenges in social media data [23]. Across multiple studies, Random Forest has consistently achieved strong performance, demonstrating reliability in scenarios where traditional models may fail [4], [8].

In addition to SVM and RF, several other machine learning algorithms have been explored for cyberbullying detection. Naïve Bayes (NB) has been applied in smaller datasets due to its efficiency, although it struggles with more nuanced linguistic contexts [7]. Logistic Regression (LR) has also been employed because of its interpretability [13], but its performance often declines in high-dimensional or complex feature spaces. Decision Tree (DT) models provide transparency in their classification process [7], yet they are prone to overfitting when applied to unbalanced datasets. Other techniques such as K-Nearest Neighbors (KNN), AdaBoost, Gradient Boosting, and Classification and Regression Trees (CART) have been tested with varying degrees of success. While some studies report promising results, their performance tends to be dataset-dependent and less consistent compared to SVM and Random Forest.

Table I presents the machine learning algorithms applied in 13 reviewed studies on cyberbullying detection. This comparison illustrates both the frequency and variety of algorithm selections, offering insights into current trends within the field.

TABLE I. COMPARISON OF REVIEWED STUDIES ON MACHINE LEARNING ALGORITHMS USED

Study	SVM	RF	NB	DT	LR	KNN	AdaBoost	Gradient Boosting	CART
[1]			✓						
[2]	✓		✓						
[3]	✓		✓						
[4]	✓	✓	✓	✓					
[5]	✓								
[6]	✓		✓						
[7]		✓	✓	✓	✓		✓		
[8]	✓	✓	✓		✓				
[9]	✓	✓	✓	✓					
[10]	✓	✓	✓		✓	✓			✓
[11]		✓					✓	✓	
[12]	✓	✓	✓	✓	✓	✓			
[13]	✓	✓	✓		✓	✓			

D. Hyperparameter Optimization

Hyperparameter optimization refers to the process of adjusting hyperparameter settings to achieve optimal model performance, and it is considered a fundamental step in machine learning [33]. This process is critical in improving the effectiveness of models designed to detect cyberbullying. If hyperparameters are poorly tuned, models risk underfitting or overfitting, which limits their ability to generalize to new and unseen data. Given that cyberbullying detection often involves noisy, imbalanced, and linguistically diverse datasets, the application of appropriate optimization strategies is vital to enhance the accuracy, reliability, and overall robustness of prediction systems.

One widely used technique is Grid Search, which systematically evaluates all possible parameter combinations to identify the best configuration. Although computationally intensive, Grid Search is particularly effective in cyberbullying detection tasks, where fine-tuning is necessary to capture subtle linguistic patterns [12], [34]. Another approach is Random Search, which randomly samples parameter combinations [35]. While it is less exhaustive than Grid Search, Random Search is faster and often more practical when dealing with large and complex search spaces [33]. A more advanced strategy is Bayesian Optimization, which models the relationship between hyperparameters and performance outcomes, allowing the search process to be more efficient and guided. However, its implementation is more complex compared to other methods [36]. Collectively, these optimization techniques are essential for maximizing the effectiveness of machine learning algorithms in detecting cyberbullying.

Table II presents a comparison of three commonly employed hyperparameter optimization methods: Grid Search, Random Search, and Bayesian Optimization. This table emphasizes their main advantages and drawbacks when applied to machine learning tasks.

TABLE II. COMPARISON OF HYPERPARAMETER OPTIMIZATION TECHNIQUES

Technique	Strengths	Limitations
Grid Search	Provides a systematic and exhaustive search, ensuring thorough evaluation of parameter combinations.	Highly computationally demanding and inefficient when applied to large or high-dimensional parameter spaces.
	It guarantees the identification of the optimal configuration.	
Random Search	Able to identify near-optimal solutions with significantly fewer evaluations.	It may overlook the exact optimal values, and outcomes can vary between runs.
	It scales well to large or complex search spaces.	
Bayesian Optimization	Prioritizes exploration of the most promising regions of the parameter space, leading to faster convergence.	More complex to implement and often requires additional fine-tuning of the optimization process itself.

E. Comparative Review of Studies

A comprehensive review of prior research on cyberbullying detection has been conducted, with particular attention to the methodologies, datasets, and performance outcomes of existing models. This review explores the prominent studies that have applied machine learning techniques to identify harmful content in social media posts.

Over the past five years, a substantial body of research has emerged, employing a variety of datasets, features, and classifiers for cyberbullying detection using machine learning techniques. Early research on Instagram comments by [1] demonstrated the value of combining BoW and lexicon-based features with Naïve Bayes, achieving accuracy rates above 87%. Although this study relied on a relatively small dataset of 350 comments, it highlighted the potential of hybrid feature engineering approaches. Next, [2] used a larger dataset of 5,628 tweets to explore sentiment analysis combined with n-grams. The study showed that SVM significantly outperformed Naïve Bayes, with SVM achieving 92% accuracy when using 4-grams. This highlighted the importance of capturing longer textual contexts for detecting nuanced abusive language. In the same year, [3] applied TF-IDF and sentiment analysis on Twitter data, finding that SVM, accuracy of 71.25%, again outperformed Naïve Bayes across all metrics. However, the study's unconventional 45-55% train-test split and limited methodological details raised concerns about reproducibility and robustness.

Expanding the scope, [4] investigated multiple algorithms, Naïve Bayes, SVM, Decision Tree, and Random Forest on Facebook and Twitter data, comparing BoW and TF-IDF. Their findings confirmed that TF-IDF features consistently enhanced performance and that SVM delivered the best results. Nevertheless, the absence of key evaluation metrics, such as precision and recall, limited the conclusions. In contrast, [5] focused on the detection of sarcasm and used NLP techniques with SVM, incorporating TF-IDF profanity, pronoun

frequency, and sentiment analysis. Their model achieved 74.50% accuracy, though the authors noted challenges in effectively handling sarcastic expressions. Around the same time, [6] also examined SVM and Naïve Bayes on a larger dataset of 20,001 tweets, finding that SVM consistently outperformed Naïve Bayes with an accuracy of 0.89. However, the absence of cross-validation and hyperparameter tuning weakened the generalizability of their results.

Subsequent studies have explored ensemble approaches and optimization methods to enhance performance. [7] compared classic algorithms like Naïve Bayes, Logistic Regression, and Decision Tree with ensemble models such as AdaBoost and Random Forest on 20,001 tweets. Random Forest achieved the best result with an accuracy of 0.92. Similarly, [8] tested multiple classifiers on four different datasets from Formspring and Twitter, showing that SVM consistently performed better, with an average accuracy of 79.3%. These results suggest that both SVM and Random Forest remain highly reliable for text-based cyberbullying detection. [9] further enhanced this by demonstrating SVM's strong performance on semantic features, with an accuracy of 0.865. However, methodological gaps in feature specification and preprocessing limit the study's reproducibility.

More recent works have leveraged larger and balanced datasets, as well as advanced optimization. [10] used over 40,000 tweets across five cyberbullying categories, applying six classifiers. Their studies highlighted the strengths of Random Forest, which outperformed all others with an accuracy of 0.941. [11] analyzed 47,000 tweets using Random Forest, AdaBoost, and Gradient Boosting. They demonstrated the effectiveness of Random Search hyperparameter tuning, which boosted Random Forest accuracy to 94%. These findings highlighted the growing role of optimization in improving model reliability. [12] proposed HYVADSVM, a hybrid approach combining VADER sentiment with SVM, and employed Grid Search for hyperparameter tuning. Tested on datasets from Twitter, Instagram, and YouTube, the model achieved near-perfect accuracy (98.83%), although severe class imbalance remained a concern. Finally, [13] conducted one of the most comprehensive studies using a global dataset of nearly 40,000 tweets, applying the Synthetic Minority Over-Sampling Technique (SMOTE) to address class imbalance and Grid Search for tuning. Random Forest again achieved the best performance, with an accuracy of 94.56%.

These studies demonstrate consistent trends in cyberbullying detection research. SVM and Random Forest repeatedly emerge as the most effective classifiers across diverse datasets, while feature engineering evolved from BoW and lexicons to TF-IDF, sentiment, and semantic features. Furthermore, recent works emphasize the importance of hyperparameter tuning, with both Grid Search and Random Search significantly enhancing model performance. Despite methodological differences, the comparative analysis underscores that hybrid approaches that integrate content-based and sentiment-based features, combined with optimized classifiers, represent the current state of the art in cyberbullying detection. A detailed comparison of significant studies is presented in Table III.

TABLE III. COMPARATIVE SUMMARY OF SELECTED STUDIES IN CYBERBULLYING DETECTION

Study	Dataset	Features	Algorithms	Optimization	Best Model	Performance
[1]	Instagram	BoW, Lexicon	NB	N/A	NB	0.872
[2]	Twitter	n-grams, Sentiment-based	SVM, NB	N/A	SVM	0.92
[3]	Twitter	TF-IDF, sentiment analysis	SVM, NB	N/A	SVM	0.713
[4]	Twitter, Facebook	BoW, TF-IDF	NB, SVM, DT, RF	N/A	SVM	N/A
[5]	Twitter	TF-IDF, Profanity + Pronouns, Sentiment-	SVM	N/A	SVM	0.745
[6]	Twitter	TF-IDF	SVM, NB	N/A	SVM	0.89
[7]	Twitter	TF-IDF	NB, LR, AdaBoost, RF	N/A	RF	0.92
[8]	Formspring, Twitter	Profanity, Sentiment-based	RF, NB, SVM, LR	N/A	SVM	0.793
[9]	Kaggle	Semantic analysis	SVM, RF, NB, DT	N/A	SVM	0.865
[10]	Twitter	TF-IDF	LR, KNN, RF, CART, NB, SVM	N/A	RF	0.941
[11]	Twitter	TF-IDF	RF, AdaBoost, Gradient Boosting	Random Search	RF	0.941
[12]	Twitter, Instagram,	TF-IDF, Sentiment-based	SVM, KNN, NB, LR, DT, RF	Grid Search	SVM	0.988
[13]	Twitter	TF-IDF	RF, SVM, LR, NB, KNN	Grid Search	RF	0.946

These studies demonstrate consistent progress in cyberbullying detection, with SVM and Random Forest emerging as reliable classifiers, particularly when enhanced through hyperparameter optimization. However, a closer examination reveals several persistent challenges and methodological gaps that limit the real-world applicability of these systems, which are discussed in detail in Section 4.

IV. LIMITATIONS AND RESEARCH GAPS

Despite significant advancements that have been made in cyberbullying detection research, several important limitations persist. One key challenge lies in the linguistic complexity of online communication. Social media discourse is often characterized by sarcasm, irony, emojis, abbreviations, and platform-specific slang, all of which can obscure the true intent of a message. As a result, many machine learning models struggle to accurately classify such ambiguous content, leading to misclassification or reduced overall accuracy [25].

Another limitation is the issue of class imbalance in available datasets. Cyberbullying instances typically represent a much smaller proportion of the data compared to non-bullying posts. This imbalance leads to models that are biased toward predicting the majority class, thereby underperforming when detecting minority categories of abuse, such as racial, religious, or gender-based harassment. While some studies have attempted to address this through techniques like SMOTE [13], many reviewed studies fail to acknowledge or systematically mitigate this fundamental challenge. The lack of balanced data not only impacts detection accuracy but also reduces the fairness and inclusiveness of these systems [13]. Models trained on imbalanced datasets may inadvertently perpetuate biases, disproportionately failing to protect vulnerable groups who are most likely to experience specific forms of targeted harassment. Furthermore, the severe imbalance reported in some studies, such as [12], documenting 98% bullying with 2% non-bullying classification, raises questions about the ecological validity of such datasets and whether they accurately represent real-world social media environments.

A further concern relates to the generalizability of trained models. Many studies rely on datasets collected from a single platform, such as Twitter, yet the linguistic style and interaction patterns on other platforms like Instagram, TikTok, or Facebook may differ considerably. Instagram emphasizes visual content with accompanying captions and comments, often centering on appearance-based harassment, while TikTok's short-form video format enables different forms of viral humiliation through features like duets and stitches. Consequently, models trained on one platform often fail to transfer effectively to others, limiting their practical applicability across diverse social media environments [23].

Lastly, the majority of existing studies remain confined to English-language datasets, creating a substantial limitation for multilingual and multicultural contexts where cyberbullying manifests differently [18]. This English-centric bias neglects the global nature of social media and the diverse linguistic communities that experience online harassment. Cross-lingual detection approaches are critical for addressing multilingual environments and for expanding the applicability of detection systems beyond English-speaking populations, which remain largely unexplored in the reviewed literature.

V. CONCLUSION AND FUTURE SCOPE

Cyberbullying continues to pose a significant threat to the safety and psychological well-being of social media users worldwide. This review has systematically analyzed 13 recent studies (2020-2025) examining machine learning approaches for automated cyberbullying detection. Despite the challenges of linguistic complexity, data imbalance, and platform-specific variations, machine learning has shown strong potential for automating the detection of harmful online content. In particular, SVM and Random Forest have consistently emerged as effective classifiers when applied with robust feature engineering and optimized using systematic hyperparameter tuning such as Grid Search and Random Search. These approaches have demonstrated the ability to capture nuanced patterns in textual data, achieving detection accuracies exceeding 94% in optimal configurations.

The review reveals several key findings, which content-based features, particularly TF-IDF combined with profanity indicators and pronoun usage, provide the most reliable detection signals. Next, hyperparameter optimization significantly improves model performance, with studies reporting accuracy gains of 6-8% after tuning, and lastly, despite high reported accuracies, critical gaps remain in linguistic complexity, class imbalance, cross-platform generalizability, and multilingual support.

Looking forward, several directions can strengthen and extend research on automated cyberbullying detection. Deep learning methods, particularly transformer-based models such as BERT or RoBERTa, offer context-sensitive analysis that can improve the handling of subtle or implicit abusive behaviors. Multimodal detection, which integrates text, images, and video, represents another promising direction for capturing cyberbullying in diverse formats commonly found on social media platforms. Notably, the majority of existing studies remain confined to a single language. Therefore, cross-lingual detection approaches are critical for addressing multilingual environments and for expanding the applicability of detection systems beyond English-language datasets. Finally, embedding AI tools within broader educational and policy-driven anti-bullying initiatives is vital to ensure that technological advancements yield meaningful real-world impact.

These directions highlight the need for continued interdisciplinary research, combining technical innovation with ethical and social responsibility, to develop comprehensive solutions for combating cyberbullying in the digital age.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of Universiti Teknologi Malaysia and the Faculty of Artificial Intelligence, UTM, Kuala Lumpur, Malaysia, for providing access to facilities, resources and support for funding this research through the Potential Academic Staff Grant (Reference No.: PY/2024/02200, Cost Center No.: Q.K130000.2757.03K71).

CONFLICT OF INTEREST STATEMENT

The authors declare that this research was conducted in the absence of any commercial or personal relationships that could be construed as a potential conflict of interest.

AUTHOR'S CONTRIBUTION

This study was fully conceptualized, designed, and executed by Che Syabila Che Shamsul. All stages of the research, including methodology development, literature review, formal analysis, and writing, were carried out by the author. The research was conducted under the supervision of Dr. Wan Noor Hamiza Wan Ali, who provided academic guidance and feedback throughout the study

REFERENCES

[1] Adikara, P. P., Adinugroho, S., & Insani, S. (2020). Detection of cyber harassment (cyberbullying) on Instagram using naïve Bayes classifier with bag of words and lexicon based features. In *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology* (pp. 64–68). https://doi.org/10.1145/3427423.3427436

[2] Atoum, J. O. (2020). Cyberbullying detection through sentiment analysis. In *2021 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 292–297). https://doi.org/10.1109/CSCI51800.2020.00056

[3] Dalvi, R. R., Chavan, S. B., & Halbe, A. (2020). Detecting Twitter cyberbullying using machine learning. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 297–301). https://doi.org/10.1109/ICICCS48265.2020.9120893

[4] Islam, M. M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., & Acharjee, U. K. (2020). Cyberbullying detection on social networks using machine learning approaches. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (pp. 1–6). https://doi.org/10.1109/CSDE50874.2020.9411601

[5] Perera, A., & Fernando, P. (2021). Accurate cyberbullying detection and prevention on social media. *Procedia Computer Science, 181*, 605–611. https://doi.org/10.1016/j.procs.2021.01.207

[6] Mahesh, K., Gothane, S., Toshniwal, A., Nagarale, V., & Gopu, H. (2021). Cyber bullying detection on social media using machine learning. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, 410–416. https://doi.org/10.32628/CSEIT217381

[7] Agrawal, T., & Chakravarthy, V. D. (2022). Cyberbullying detection and hate speech identification using machine learning techniques. In *2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS)* (pp. 182–187). https://doi.org/10.1109/ICPS55917.2022.00041

[8] Ali, A., & Syed, A. M. (2022). Cyberbullying detection using machine learning. *Pakistan Journal of Engineering and Technology, 3*(2), 45–50. https://doi.org/10.51846/vol3iss2pp45-50

[9] Siddhartha, K., Kumar, K. R., Varma, K. J., Amogh, M., & Samson, M. (2022). Cyber bullying detection using machine learning. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)* (pp. 1–4). https://doi.org/10.1109/ASIANCON55314.2022.9909201

[10] Balet, T., Vo, Q., Salem, O., & Mehaoua, A. (2023). Cyberbullying detection on tweets from Twitter using machine learning algorithms. In *2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)* (pp. 177–182). https://doi.org/10.1109/ICCNS58795.2023.10193450

[11] Mathur, S. A., Isarka, S., Dharmasivam, B., & D, J. C. (2023). Analysis of tweets for cyberbullying detection. In *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*. https://doi.org/10.1109/ICSCCC58608.2023.10176416

[12] Ernawati, S., Frieyadie, F., & Yulia, E. R. (2024). HYVADSVM: Hybrid VADER-SVM and GridSearchCV optimization for enhancing cyberbullying detection. *Journal of Robotics and Control (JRC), 6*(1), 101–113. https://doi.org/10.18196/jrc.v6i1.24385

[13] Sayed, F. R., Elnashar, E. H., & Omara, F. A. (2025). Cyberbullying detection in social media using natural language processing. *Scientific African*, e02713. https://doi.org/10.1016/j.sciaf.2025.e02713

[14] Ariffin, A., Mohd, N., & Rokanatnam, T. (2021). Cyberbullying via social media: Case studies in Malaysia. *OIC-CERT Journal of Cyber Security, 3*(1), 21–30.

[15] Giumetti, G. W., & Kowalski, R. M. (2022). Cyberbullying via social media and wellbeing. *Current Opinion in Psychology, 45*, 101314. https://doi.org/10.1016/j.copsyc.2022.101314

[16] Zhu, C., Huang, S., Evans, R., & Zhang, W. (2021). Cyberbullying among adolescents and children: Comprehensive review of the global situation, risk factors, and preventive measures. *Frontiers in Public Health, 9*. https://doi.org/10.3389/fpubh.2021.634909

[17] Saruddin, M. Z., Azman, A. Z. F., Mohd Zulkefli, N. A., & Isa, M. R. (2025). Prevalence of cyberbullying and its associated factors among high school students in Klang District, Malaysia. *Malaysian Journal of Medicine and Health Sciences, 21*(3), 220–231. https://doi.org/10.47836/mjmhs.21.3.26

[18] Khairy, M., Mahmoud, T. M., & Abd-El-Hafeez, T. (2021). Automatic detection of cyberbullying and abusive language in Arabic content on social networks: A survey. *Procedia Computer Science, 189*, 156–166. https://doi.org/10.1016/j.procs.2021.05.080

[19] Gutiérrez-Esparza, G. O., Vallejo-Allende, M., & Hernández-Torruco, J. (2019). Classification of cyber-aggression cases applying machine learning. *Applied Sciences, 9*(9), 1828. https://doi.org/10.3390/app9091828

[20] Shapiro, G. L. (2022). Cyberbullying: Introduction, definition, and impact on mental health in youth. *Journal of the American Academy of Child & Adolescent Psychiatry, 61*(10), S71. https://doi.org/10.1016/j.jaac.2022.07.299

[21] Dhamodharan, M., & Sunaina, K. (2023). Cyberbullying. In *Advances in Human and Social Aspects of Technology* (pp. 224–249). https://doi.org/10.4018/978-1-6684-5760-3.ch010

[22] Kartik. (2023). Cyber bullying. *International Journal for Multidisciplinary Research, 5*(5). https://doi.org/10.36948/IJFMR.2023.V05I05.6417

[23] Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., Mujtaba, G., Chiroma, H., Khattak, H. A., & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. *IEEE Access, 7*, 70701–70718. https://doi.org/10.1109/ACCESS.2019.2918354

[24] Wang, A., & Potika, K. (2021). Cyberbullying classification based on social network analysis. In *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 87–95). https://doi.org/10.1109/BigDataService52369.2021.00016

[25] Mahdi, M. A., Fati, S. M., Hazber, M. A., Ahamad, S., & Saad, S. A. (2024). Enhancing Arabic cyberbullying detection with end-to-end transformer model. *Computer Modeling in Engineering & Sciences, 141*(2), 1651–1671. https://doi.org/10.32604/CMES.2024.052291

[26] Kumar, K. M. K., Ullas, K., Reddy, V. S., Snehitha, M. S., Malkhed, M. E., & Kumar, K. D. (2024). Detection of bullying text: A multi-faceted approach using machine learning and natural language processing. In *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 180–186). https://doi.org/10.1109/ICOSEC61587.2024.10722479

[27] Woo, W. H., Chua, H. N., & Gan, M. F. (2023). Cyberbullying conceptualization, characterization and detection in social media: A systematic literature review. *International Journal on Perceptive and Cognitive Computing, 9*(1), 101–121. https://doi.org/10.31436/IJPCC.V9I1.374

[28] 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 social media text. *PLoS ONE, 13*(10), e0203794. https://doi.org/10.1371/journal.pone.0203794

[29] Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing, 11*(1), 3–24. https://doi.org/10.1109/TAFFC.2017.2761757

[30] Al-Hashedi, M., Soon, L., Goh, H., Lim, A. H. L., & Siew, E. (2023). Cyberbullying detection based on emotion. *IEEE Access, 11*, 53907–53918. https://doi.org/10.1109/ACCESS.2023.3280556

[31] Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J. P., Kowalski, R., Hu, H., Luo, F., Macbeth, J., & Dillon, E. (2016). Cyberbullying detection with a pronunciation-based convolutional neural network. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 740–745). https://doi.org/10.1109/ICMLA.2016.0132

[32] Thun, L. J., Teh, P. L., & Cheng, C. (2021). CyberAid: Are your children safe from cyberbullying? *Journal of King Saud University – Computer and Information Sciences, 34*(7), 4099–4108. https://doi.org/10.1016/j.jksuci.2021.03.001

[33] Monica, & Agrawal, P. (2024). A survey on hyperparameter optimization of machine learning models. In *2024 2nd International Conference on Disruptive Technologies (ICDT)* (pp. 11–15). https://doi.org/10.1109/ICDT61202.2024.10489732

[34] Winarno, W., Wiranto, W., & Harjito, B. (2023). Enhancing machine learning performance in cyberbullying detection through hyperparameter optimization. In *2023 International Conference on Technology, Engineering, and Computing Applications (ICTECA)* (pp. 1–6). https://doi.org/10.1109/ICTECA60133.2023.10490843

[35] Dabool, H., Alashwal, H., Alnuaimi, H., Alhouqani, A., Alkaabi, S., & Ahbabi, A. A. (2024). Comparative analysis of hyperparameter tuning methods in classification models for ensemble learning. In *2024 7th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI)* (pp. 1–5). https://doi.org/10.1109/ACAI63924.2024.10899492

[36] Alibrahim, H., & Ludwig, S. A. (2021). Hyperparameter optimization: Comparing genetic algorithm against grid search and Bayesian optimization. In *2022 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1551–1559). https://doi.org/10.1109/CEC45853.2021.9504761

[37] Wang, S., Zhu, X., Ding, W., & Yengejeh, A. A. (2022). Cyberbullying and cyberviolence detection: A triangular user–activity–content view. *IEEE/CAA Journal of Automatica Sinica, 9*(8), 1384–1405. https://doi.org/10.1109/JAS.2022.105740.