Artificial Intelligence Techniques for Cyberbullying Detection: A Comprehensive Review of Machine Learning and Deep Learning models

Abstract. Cyberbullying has emerged as a critical threat in online social platforms, particularly among adolescents and vulnerable users. The rapid growth of internet and social media user's generated content on social media has intensified the need for automated systems capable of detecting harmful content. This survey provides a comprehensive overview of recent advancements in cyberbullying detection using traditional machine learning (ML)methods such as SVM, BN and AdaBoost, deep learning (DL) models like CNN and transformers, and hybrid approaches that combine ML and DL models. We review a wide range of studies published between 2023 and 2025, highlighting the evolution of used techniques. Traditional ML methods have given high performance for text classification. However, DL models like CNNs, LSTMs and transformers have significantly improved performance by capturing contextual and semantic patterns in textual data. Moreover, hybrid systems that integrate ML and DL are increasingly being adopted to combine the strengths of both techniques. The survey also discusses datasets used for training and evaluating models. This work aims to guide future research for more pertinent and robust solutions for cyberbullying detection.

Keywords: cyberbullying, machine learning, deep learning, transformers, Artificial intelligence

1 Introduction

The rapid increase of internet and social media users has transformed global communication and connected billions of individuals throughout the world. However, the augmented connectivity presents many negative phenomena, among which is the exponential rise of undesirable online contents. Of these phenomena, cyberbullying has emerged as one of the most risky phenomenon. Cyberbullying is defined as repeated intentional aggressive behavior delivered through electronic or digital means, and can manifest itself in many forms including textual harassment, hate messages, cyberstalking and propagation of harmful visual content such as videos and images [1]. Different from traditional bullying, cyberbullying can reach international audiences, making its impact particularly harmful. Victims of cyberbullying often suffer from grave mental anxiety and disturbances such as decreased self-esteem and social isolation. Additionally, cyberbullying can also lead to self-harming or suicidal inclinations.

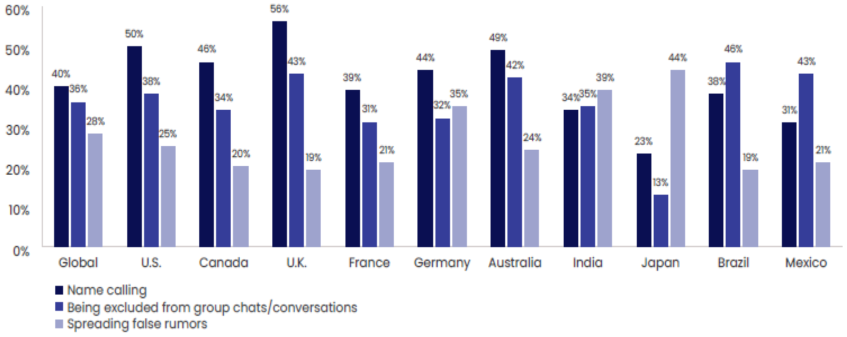

Figure 1 cyberbullying forms [2]

The precedent bar graph (cyberbullying formsFigure 1) presents frequencies among different countries of various cyberbullying forms. Globally, the most common form is name-calling, found among 40% participants. In US, UK, France, Germany, Australia, and Brazil, likewise, name-calling is also found as the most common behavior. On the other hand, Japan's most common cyberbullying form is being excluded from group chats or conversations at 44%, a much larger proportion than other regions.

This significant and profound impacts cyberbullying underline the need for effective detection approaches. Consequently, researchers used advanced techniques for cyberbullying detection, among these techniques, artificial intelligence approaches including machine learning (ML) and deep learning (DL).

The objective of this article is to provide a review of the latest research utilizing artificial advanced intelligence techniques such as machine learning and deep learning for cyberbullying detection. The next part of this study contain the following sections: section 2: Related Work where we present an overview of the evolution approaches used for cyberbullying detection. Section 3: Cyberbullying Detection Using AI: in this section, we discuss recent studies that have used AI approaches. This section is divided into three subsections; the first one presents traditional ML methods used for cyberbullying detection with the advantages and disadvantages of these methods. In the second subsection, we discuss DL models and their significant impact in cyberbullying detection. In the last subsection we present hybrid approaches that combine several DL and ML models. We conclude this survey with a conclusion that summaries cited approaches and our research perspectives for the future.

2 Related work

Although researches on cyberbullying detection dates back to 2000 **[3]**. At this period, the developed systems were based on rule-based approaches. These approaches employed lexical detection techniques based on keyword lists and linguistic rules in order to identify hateful and offensive content.

With the notable increase of cyberbullying on internet, the number of studies on cyberbullying detection has been also observed since 2010. This growth coincided with the emergence of the first researches using artificial intelligence methods for cyberbullying detection. Over the last ten years, researches in this area have experienced exponential rate increase, particularly with the use of deep learning models that have significantly enriched detection approaches. The bar graph below (Figure 2), show the number of studies on cyberbullying detection per year.

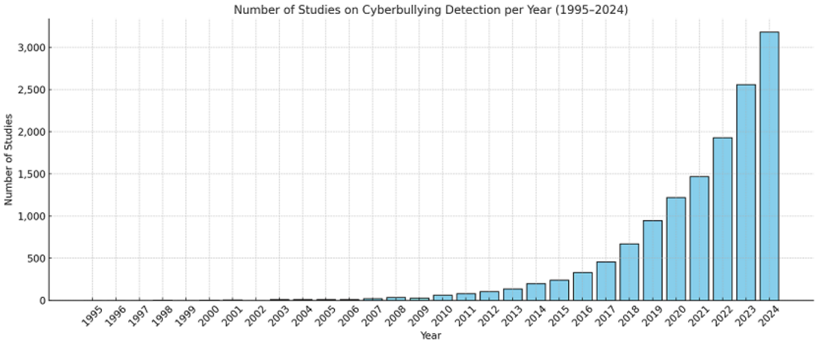

Figure 3 Researches on Cyberbullying detection from 1995 to 2024

In recent years the most used AI technique for cyberbullying detection are ML methods and DL models, while some studies propose hybrid models that combine ML and DL to develop more powerful systems. The table below present the most AI techniques used for cyberbullying detection classified in three categories: Traditional ML, DL and hybrid models with the main role of each category.

Table 1 AI techniques used for cyberbullying detection

Model Type	Primary Role	Examples
Traditional ML	Classification	SVM, Naive Bayes, AdaBoost (TF-IDF/Word2Vec for features extraction).
Deep Learning	Feature Extraction + Classification	BERT (embeddings + classification),
Hybrid Models	DL for features extraction + ML for classification	CNN-RF, ArabicBERT-BiLSTM-RBF.

3 Cyberbullying detection using AI

3.1 Traditional machine learning for Cyberbullying detection

Traditional machine learning algorithms such as: SVM; NB, LR and RF, are widely used in cyberbullying detection. These methods have given strong results in several researches, including SVM, NB and AdaBoost.

Support Vector Machine (SVM): The SVM is one of the most used traditional machine learning algorithm, spicily in cyberbullying detection. For instance, in [4] implemented six supervised ML algorithms to classified tweets into three classes: hate speech, offensive language, and normal text. The SVM algorithm achieves high accuracy (90.98%). Similarly**,** in[6] several classifiers were employed to detect cyberbullying in social media platforms, the results showed (SVM) as the most effective algorithm with an accuracy rate of 90.06%. In addition**,** in [10] the SVM algorithm was used with a set of machine learning algorithms and deep learning models to develop an automated system to detect cyberbullying in Turkish dataset, the developed model reached significant results with 82% F1 score. Furthermore, in [23] seven machine learning classifiers were implemented and evaluated including SVM that reach an accuracy of 95.74%. In [5], multiple traditional machine learning were used for automatic cyberbullying detection, these models were evaluated using a global dataset of 37,737 tweets. However, the SMV algorithm didn't perform well in [5] and gave the lowest score with accuracy of 67.13%.

These studies showed that the SVM algorithm is more effective for linear and nonlinear separation; however it is less powerful with large datasets as in [5] where the SMV gave a low accuracy of (67.13%).

Naive Bayes (NB): The NB algorithm is employed in multiple studies [4, 7, 9] for its simplicity and speed, particularly multinomial NB that was used for text classification. For instance, in [9] various supervised machine learning techniques were implemented to detect cyberbullying in Urdu texts. The BN was used for multi-class classification and it delivered good performance with an accuracy score of 91.87%. Moreover, in [4] used using supervised machine learning (ML) techniques to detect cyberbullying in Twitter dataset. The NB algorithm reached a high accuracy score of 88.82% accuracy and performing well in detecting offensive language; however the algorithm failed to detect hate speech. Similarly, in [7] the NB algorithm was utilized with other ML algorithm to develop a reliable machine-learning model for social media cyberbullying detection in the Bengali language. All the used ML algorithms gave almost similar accuracy scores ranging from 76% to 79%. However, the NB algorithm achieved the lowest accuracy score about 76%.

AdaBoost: The AdaBoost (ADB) algorithm is commonly used in cyberbullying detection studies. For example, in [12] authors proposed a model for detecting cyberbullying using multiple ML algorithms, the ADB algorithm in the proposed model achieved a high rate of 86.52% accuracy. Similarly, in [5], the researchers implement and evaluate seven supervised ML classifiers, the ADB algorithm recorded strong results with 89.30% accuracy and an F1-score of 91.66%. The results demonstrate that ADB classifier was effective for improving weak learners but it still less common compared to other traditional models such as SVM or NB.

Other traditional machine learning algorithms: In multiples studies we find that other traditional machine learning algorithms were used to develop automatic system for detecting cyberbullying. Algorithms such as: Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), XGBoost, Stochastic Gradient Descent (SGD), and Extra Tree Classifier have showed important results, LR and RF showing strong performance (LR Recorded 90.03% accuracy in [5] and RF achieved the highest performance with 91.08% accuracy in [17]).

Table 2 Traditional ML Model in cyberbullying detection

Study	year	Traditional ML Model	Feature Extraction	Dataset	Accuracy
[6]	2024	SVM	TF-IDF, Word2Vec	Social media platforms (WhatsApp, Facebook, Instagram, TikTok, YouTube)	90.06%
[4]	2024	SVM, RandomForest Decision Tree Naive Bayes Logistic Regression KNN	BoW, TF-IDF	Twitter (24,783 tweets)	SVM: 90.98% LR: 90.03% RF: 89.28% NB: 88.82% DT: 79.40% KNN: 80.87%
[5]	2020	LR, LGBM, SGD,RF , AdaBoost, Multinomial NB, SVM	TF-IDF, Word2Vec	Twitter (37,373 tweets)	LR: 90.57%
[7]	2023	Naive Bayes, SVM,	TF-IDF	Facebook (44,001 comments)	SVM: 79% NB: 76%
[12]	2024	AdaBoost,	CountVectorizer	Unspecified social media (likely Twitter/YouTube)	AdaBoost: 86.52%
[17]	2023	Random Forest AdaBoost, Multinomial NB, SVM	TF-IDF	Facebook (11,000 Bangla comments)	RF: 91.08% Multinomial-NB: 89.44%

Traditional ML algorithms are widely used for cyberbullying detection. These algorithms demonstrate their effectiveness and provide strong results in many studies. For instance, they provide high performance with structured data and were more effective with smaller Datasets [10]. In Addition, traditional ML methods are easy to implement and interpret and require less expertise for model design and implementation [4 , 20]. Moreover, the results of traditional ML methods are easy to interpret, which facilitates the decision-making process [5 , 12]. Furthermore, the combination of traditional ML methods with feature extraction techniques such as: TF-IDF, Word2Vec and CountVectorizer have yielded significant scores in several studies [4 **,** 8]. However, one of the most significant disadvantages of traditional ML methods is their inability to capture sarcasm, irony and culture specific language. This problem has been mentioned in several studies such as [6 ,11 , 29] .Also, traditional ML methods are sensitive to unbalanced classes which are widely used in training datasets for cyberbullying detection models, which decreases the performance of the proposed models [4, 8 , 12] and some of them have had problems when used with large datasets and require more computational resources.

3.2 Deep learning models for Cyberbullying detection

Multiple deep learning models have been employed for cyberbullying detection including CNN, LSTM and transformers.

Convolutional Neural Networks (CNN) is one of the most deep learning model utilized in for cyberbullying detection. For instance, [8] authors used several deep learning models to design and implement a new framework accommodating including CNN, the proposed model provide high classification results by achieving a score of 95.6 % in accuracy. In addition, in [10] researchers employed multiple classification techniques, including CNN, the result showed that the CNN model outperformed other deep learning models; it achieved an average F-score of 92.11%. Similarly, in[14] developed a deep learning model to identify cyberbullying in Turkish Twitter posts. The proposed CNN model achieved (81.6%) in accuracy, In [32] present an effective Arabic cyberbullying detection system, AraCB, using deep learning models, specifically convolutional neural networks (CNNs). The results showed a rate of 16.5% accuracy improvement. In [23] implemented effective model for detecting cyberbullying in Arabic social media content, several deep learning models were used and CNN showed a rate of accuracy (92%),

Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models were applied in several studies. For example, in [8] researchers proposed hybrid model for detecting cyberbullying based on some deep learning algorithms. The BiLSTM models performed best among used neural networks and reached a high accuracy score of 96.88 %. Additionally, in [25] the authors proposed a system for detecting Arabic hate speech and offensive language on social networks, using the Bidirectional long-term memory (Bi-LSTM) model, they achieved strong results with an accuracy of 96.35% and an F1-score of 85.82%. Also, in [20] Arabic cyberbullying detection by proposing hybrid models to identify and classify abusive content on social media, results achieved by Deep learning models (BiLSTM 94.17%, LSTM 93.10%) showed that BiLSTM outperforms LSTM by processing text bidirectionally. Furthermore, in [10] the researchers employed several classification techniques to detect cyberbullying in Turkish text, including LSTM and BiLSTM techniques. The two models achieved good and close results. LSTM model achieved an F1-score of 87.83 % and the BiLSTM model achieved an F1-score of 87.87%. Similarly, In [30] they developed a robust detection system using multiple machine learning and deep learning techniques such as LSTM and BiLSTM. Results achieved by DL models were important especially BiLSTM (F1: 0.74, accuracy: 74.59%) and LSTM (F1: 0.74, accuracy: 74.66%).

The LSTM and BiLSTM models are commonly used for cyberbullying detection; these models demonstrate their effectiveness with languages hard to process like Arabic [20, 25] and Turkish[10].

Transformers are among the most common deep learning techniques used for cyberbullying detection and BERT is the most used transformer in the examined studies for this review. For example, in [11] the authors propose a model that leverages a range of features for identifying cyberbullying, focusing on the bidirectional deep learning model BERT. The proposed BERT-based model gave strong results by achieving an accuracy score of 91.90%. Additionally, in [18] used LLM for detecting cyberbullying in Bengali social media texts, the study employed two transformer-based models (BanglaBERT and mBERT) for text classification. The BanglaBERT achieved the highest accuracy of 88.04%, with an F1 score of 87%. Similarly, in [21] authors used a set of LLMs such as : Fine-tuned RoBERTa, BERT, XLNet, and XLM-RoBERTa, for automated cyberbullying detection on social media platforms, RoBERTa model achieved the best performance with a F1-score of 66% for the bullying class, followed by BERT (F1-score of 64%), XLM-RoBERTa (F1-score of 60%), and XLNet (F1-score of 52%). In [30] XLM-RoBERTa model was also employed to classify cyberbullying texts into five categories: Cy-Flaming, Cy-Threat, Cy-Racism, Cy-Pull-a-Pig, and Not Bullying. The developed system based on XLM-RoBERTa achieved a good performance with an F1-score of 83% and accuracy of 82.61%. Furthermore, in [20] two hybrid transformer models were developed to identify and classify abusive content on social media, particularly in Twitter. The first model combined CAMeLBERT with AraGPT2, and the second model combined AraBERT with XLM-R. The CAMeLBERT + AraGPT2 model achieved the highest performance, with 97.57% accuracy, 97.43% precision, 97.51% recall, and 97.47% F1-score, outperforming AraBERT + XLM-R (95.19% accuracy) and individual transformers (e.g., CAMeLBERT 96.97%). On another study, In [19] the authors employed effectiveness of large language models (LLMs), specifically a fine-tuned GPT-3 Ada model, for detecting cyberbullying on social media platforms, the results showed that GPT-3 Ada achieved high results with an accuracy score of 90%, precision 92%, recall 88%, and 93% F1 score. The GPT-3 model achieved similar results and high accuracy in other studies like in studies [11] and [23] where it achieved 91.90% and 98.45% respectively.

Other deep learning model such as Gated Recurrent Unit (GRU), Bidirectional GRU (BiGRU), Attention-BiLSTM, and hybrid CNN-BiLSTM were used in studies

[8] for cyberbullying detection. Also, Multi-Layer Perceptron (MLP) was employed by[14] for Turkish text classification, the proposed model achieved promising results with an accuracy score of 93.2%, but these models are not commonly used in reviewed studies.

The table below summaries studies that utilized DL models for cybirbullying detection, with the data set used in each studies.

Table 3 Deep learning models on cyberbullying detection

Study	year	DL Model	Feature Extraction	Dataset	accuracy
[10]	2024	CNN	TF-IDF, GloVe	Turkish offensive comments	CNN:92.11 %
[14]	2024	MLP	Word embeddings, social media metadata	Turkish Twitter (5,000 posts)	MLP: 93.2%
[18]	2024	LSTM, GRU, Bangla-BERT, mBERT	TF-IDF, Word2Vec	Facebook, YouTube (10,000 Bengali comments)	BanglaBERT: 88.04%,
[23]	2024	E-BERT, CNN,Bi-GRU, LSTM, Bi-LSTM	WordPiece tokenization	Arabic tweets (X platform)	E-BERT: 98.45% CNN: 92% Bi-GRU: 93% LSTM: 83.18% Bi-LSTM: 82.12%
[30]	2024	GRU, CNN, LSTM, BiLSTM, m-BERT, Bangla-BERT, XLM-RoBERTa	TF-IDF, BoW, Keras embeddings	CBD (2,751 Bengali texts)	XLM-RoBERTa: 82.61% Bangla-BERT: 81.16% m-BERT:76.66% BiLSTM:74.59 % LSTM:74.66% GRU:72.73% CNN: 70.66%
[19]	2023	GPT-3 Ada, BERT	Not specified	Twitter (47,000 tweets)	GPT-3 Ada: 90% BERT: 91%
[21]	2023	RoBERTa, BERT, XLNet, XLM-	TF-IDF, SBERT	Formspring (D1:11,997 non-bullying, 776bullying),	RoBERTa: 66% F1 (D1), 87% F1 (D2);

		RoBERTa		(D2:19,533 bullying, 19,526 non-bullying)	
[22]	2023	LSTM, BiLSTM, BERT,RoBERTa, GPT-3	BoW, TF-IDF, GloVe, KerasE, BERT	Twitter (99,991 tweets)	BERT: 99.7% RoBERTa: 99.7%
[8]	2021	CNN,LSTM, Bi-LSTM, GRU, Bi-GRU, CNN-iLSTM, Attention-BiLSTM	GloVe, FastText, Paragram	Two real-world datasets	CNN 95.6 %

Deep learning models are widely applied for cyberbullying detection. These models demonstrate an excellent performance in capturing complex linguistic patterns and contextual nuances by achieving high accuracy in many studies. These models are very effective in capturing the semantics of various forms of cyberbullying such as nuances and sarcasm, which are harder to identify. For example, both BERT and RoBERTa models achieved high performance in [22]. Additionally, DL models performed very well and achieved excellent results with large and diverse datasets. For example, in [8], the CNN-RF hybrid model achieved 98.41% accuracy on a dataset of 100,000 tweets, and in [22], BERT model reached 99.8% F1-score on a dataset of 99,991 tweets. Moreover, they perform well with imbalanced data, contrary to traditional ML methods. Despite this, DL models require more computational resources than traditional ML methods, and in several studies we find that transformers such as BERT and RoBERTa performed less effectively under limited conditions [20, 22, 23]. In Addition, they required large and diverse datasets to achieve high performance, and performed weakly with small dataset.

3.3. Hybrid models for Cyberbullying detection

Many researchers have turned to the characteristics of both approaches to take advantage of them. They have employed DL techniques for feature extraction and traditional ML techniques for classification. Additionally, some studies developed hybrid models by combining two or more deep learning models including transformers. For instance, authors in [16] developed ProTect a hybrid deep learning model for proactive detection of cyberbullying on social media platforms. They combined CNN/LSTM for feature extraction and ML algorithms such as RF, SVM and NB for text classification, leveraging deep learning’s semantic capabilities with ML’s efficiency. The hybrid CNN-RF method performed the best and achieved high results: accuracy 98.41%, precision 99.71%, recall 94.19% and F1-score 96.87. Similarly, in[26] proposed hybrid models that consist of a combination between BERT models, deep learning models, and a traditional classifier in a cascaded manner. The cascaded model ArabicBERT-BiLSTM-RBF achieved an accuracy score of 98.4%, performing other developed models.

On the other hand**,** in [32] proposed an Arabic cyberbullying detection system called AraCB, this model integrate multiple deep learning techniques including CNN with ReLU Activation and Average Pooling for to process and classify text effectively. The AraCB model reached high accuracy score of 82.6%. In the same way, In [13] used deep learning techniques to develop several hybrid models for multilingual cyberbullying detection in Bangla and Chittagonian texts. The proposed models are: BiLSTM+GRU, CNN+LSTM, CNN+GRU, CNN+BiLSTM, and (CNN+LSTM)+ BiLSTM. The results achieved by these models were favorable, with accuracy between 80% and 86%, and the (CNN+LSTM)+BiLSTM produced the highest results among them. Also, In [33] presented Firefly-CDDL an hybrid model for cyberbullying detection in tweeter. The model combines a Convolutional Neural Network (CNN) with the Firefly Algorithm (FA) to optimize the CNN's structure. The proposed model reached strong results by achieving a high accuracy score of 98.75%. Additionally, in [34], presents a hybrid deep learning model for cyberbullying detection in Chinese social media texts. The proposed model combines the pre-trained language model XLNet, with a deep Bi-LSTM (Bidirectional Long Short-Term Memory) and was used to classify Chinese social media texts as cyberbullying or non-cyberbullying. This hybrid model gave high results and achieving 90.43% of F1-score.

Table 4 Hybrid models for cyberbullying detection

Study	year	Hybrid Model	Feature Extraction	Dataset	Accuracy
[20]	2025	CAMeLBERT +AraGPT2, AraBERT+ XLM-R	Feature fusion	Arabic Twitter (17,670 tweets)	CAMeLBERT +AraGPT2: 97.57% AraBERT+ XLM-R: 95.19%
[16]	2024	CNN/LSTM + ML classifier	Deep learning features + metadata	Instagram, Twitter (1Mcommen ts,4M users)	Not specified
[28]	2024	BERT with dual attention, hierarchical embeddings	Sentiment, topic embeddings	Not specified	91%
[32]	2024	CNN + Multi-Head Attention + ResNet	Word2Vec, TF-IDF	ArCybC (4,505 Arabic tweets)	16.5% accuracy improvement,

Hybrid approaches seek to leverage the benefits of two or more techniques. For example, some studies combine deep learning's contextual understanding with ML's computational efficiency [16, 26]. While, other studies combine multiple deep learning models like Bi-LSTM or CNN to develop more powerful models for cyberbullying detection [34, 13].

3.4. Datasets

In reviewed studies we find that researchers focus more on detecting cyberbullying on Twitter. For example, in [4] the authors proposed an automatic system for identifying abusive text. They used Kaggle dataset, containing about 24,783 tweets in English which were classified into three categories: hate speech, offensive language and normal text. Similarly, in [5]the proposed model used a global English dataset of 37,373 unique tweets from twitter. Furthermore, in [22] used a large datatset with more than 99,000 tweets. In addition, Twitter datasets were used in several studies for detecting cyberbullying in other languages such as Arabic in [32 ,20 , 23, 26], Turkish in[14] and Urdu in [9].

On the other hand, some studies used datasets collected from diverse platforms, such as Facebook, Instagram, YouTube, WhatsApp, TikTok and also Twitter. For instance, in [18]_LLM used a dataset of 10,000 comments from Bengali comments from Facebook and YouTube. Moreover, in [6] the authors used a dataset collected from several social media platforms (WhatsApp, Facebook, Instagram, TikTok, YouTube).

Indeed, the majority of datasets utilized for cyberbullying detection are in English; nevertheless, a few studies exist on datasets in Arabic [20, 23, 25, 26], Bengali [7, 18, 30], Turkish [10] and [14], Urdu [9], and code-mixed languages [31]. In Addition, studies on multi-language datasets are very rare.

The following table summarizes the different datasets used in the examined studies.

Table 5 used dataset in cyberbullying detection studies

Dataset	Language	Size	Study
Arabic tweets dataset	Arabic	4,505 to 17,670 tweets	[20] - [23] - [25] - [26] - [32]
Bangla Dataset (CBD) from Facebook, YouTube, Instagram	Bengali	44,001 2,751 texts (comments)	[7] - [17] - [18] - [30]
Code-mixed meme captions dataset	English + regional languages	Not specified	[31]
Twitter, WhatsApp, Facebook, Instagram, TikTok, YouTube, gaming platforms	English	From 12,773 to 99,991 tweets/comments	[4] - [5] - [6] - [11] - [15] - [16] -[19] - [21] - [22] [24] - [27] - [29]
Urdu tweets dataset	Urdu	7,625 tweets	[9]
Turkish Twitter posts	Turkish	5,000 tweets	[10] - [14]

Twitter is the most dominant Dataset among various source for cyberbullying detection studies with datasets ranging from 5,000 [14] to 99,000 tweets [22]. The used datasets are mostly in English, but there are some studies that work on language-specific datasets like Arabic, Turkish and Urdu.

4 Conclusion

In this study we have presented comprehensive overview of recent AI techniques used in cyberbullying detection, focusing on machine learning, deep learning, and hybrid models. Early studies have used rule-based systems for text classification. Then traditional machine learning classifiers such as SVM, BN and AdaBoost have been utilized for cyberbulying detection combined with baselines features extraction techniques like TF-IDF and BoW. The evolution of AI models like CNNs, LSTMs, and transformers led to their adoption for cyberbullying detection, because of their high performance with unstructured data. Also, hybrid approaches have emerged; these approaches combine traditional ML and DL models to improve the accuracy of cyberbullying detection.

Despite of the significant progress in this field, many challenges remain namely multilingual data and multimodal content. In this context, we see promising prospective in the use of Large Language Models (LLMs) such as BERT, RoBERTa, and GPT. Their ability to capture deep semantic relationship and contextual nuances makes these models well-suited for cyberbullying detection.

References

[1] StopBullying.gov. (n.d.). What is cyberbullying? U.S. Department of Health and Human Services. from https://www.stopbullying.gov/cyberbullying/what-is-it. visited 20/06/2025.

[2] McAfee, LLC. Cyberbullying in plain sight: 2022 global report. McAfee, LLC. https://www.mcafee.com. 20/06/2025

[3] B.Gordon Kennelly: Caught in/on the Web: To Publish Without Perishing in the Digital Age. First Monday, 5(8). https://firstmonday.org/issues/issue5_8/kennelly/index.html.

[4] C.Bhatt, P.Goyal, G.Prasad Dubey, S. Singh , V.Kumar: Detection of cyber-bullying in social-media using lassification algorithms of machine learning, Community practitioner: the journal of the Community Practitioners' & Health Visitors' Association , 21(05), p793 - 803 ,(2024)

[5] A.Muneer , S.Mohamed Fati: A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter, Future Internet,12(187) p 1-20, (2020)

[6] S.Islam, A.Noor Orno, M.Arifuzzaman: Approach to Social Media Cyberbullying and Harassment Detection Using Advanced Machine Learning . Research Square. (2024)

[7] S.Saha, S.Islam, M.Alam, M.Rahman, Z.Hasan Majumder , S.Alam and K.Hossain: Bengali Cyberbullying Detection in Social Media Using Machine Learning Algorithms , in 5th International Conference on Sustainable Technologies for Industry 5.0 (STI), IEEE, Dhaka(2023)

[8] C.Raj , A.Agarwal , G.Bharathy , B.Narayan and M.Prasad:Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques. Electronics, 10(2810).(2021)

[9] S.Khan and A.Qureshi: Cyberbullying Detection in Urdu Language Using Machine Learning, in International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering, pp. 1-6. Pakistan(2022)

[10] A.Najib, S. Ayse Ozel, O.Coban, R.Aftab Naseem Khan : Cyberbullying Detection Using Machine Learning and Deep Learning for Turkish Text , Media, Communication and Informatics Symposium. 65-80.(2024)

[11] A.Desai, S.Kalaskar, O.Kumbhar, R. Dhumal: Cyber Bullying Detection on Social Media using Machine Learning. ITM Web of Conferences. 40(03038), (2021)

[12] V.Kulkarni, S.Chakrabarti, S.Salunke, V.Wagh: A Comprehensive Analysis of Cyberbullying Detection Using Various Machine Learning Approaches. Multimedia Tools and Applications, p15-25. (2024)

[13] T.Mahmud, M.Ptaszynski, F.Masui :Exhaustive Study into Machine Learning and Deep Learning Methods for Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts. Electronics, 13(1677). (2024)

[14] Ç.Oguz Aliyeva, M.Yaganoglu: Deep learning approach to detect cyberbullying on twitter. Multimedia Tools and Applications(2024)

[15] K.Gutierrez-Batista, J.Gomez-Sanchez, C.Fernandez-Basso: Improving automatic cyberbullying detection in social network environments by fine-tuning a pre-trained sentence transformer language model. Social Network Analysis and Mining 14(136). (2024)

[16] T. Nitya Harshitha, M. Prabu1, E. Suganya, S. Sountharrajan, D.Prasad Bavirisetti, N.Gadde1, L.Sahithi Uppu : ProTect: a hybrid deep learning model for proactive detection of cyberbullying on social media . Front. Artif. Intell. 7(1269366). (2024)

[17] T.Ahmed, A.Hossain Antar, M.Rahman, A.Zafor Muhammad Touhidul Islam, D.Das, G.Rashed: Social Media Cyberbullying Detection on Political Violence from Bangla Texts Using Machine Learning Algorithm. Journal of Intelligent Learning Systems and Applications, 15, 108-122, (2023).

[18] K.Saifullah, M.Ibrahim Khan, S.Jamal, H. Sarker Iqbal: Cyberbullying Text Identification: A Deep Learning and Transformer-based Language Modeling Approach. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, 11(1), (2024) .

[19] D.Ottosson : CyberbullyingDetectiononsocialplatforms using LargeLanguageModels . Final Project, Computer Engineering BA (C), Mid Sweden University, Östersund, Sweden (2023).

[20] A. Alsuwaylimi1 Amjad, S. Alenezi Zaid: Leveraging Transformers for Detection of Arabic Cyberbullying on Social Media: Hybrid Arabic Transformers. Comput Mater Contin.;83(2) , pp 3166 -3185,(2025)

[21] B.Ogunleye, B. Dharmaraj : The Use of a Large Language Model for Cyberbullying Detection. Analytics, 2, 694–707. (2023)

[22] M.Ahmadinejad, N.Shahriar, L.Fan: Self-Training for Cyberbullying Detection: Achieving High Accuracy with a Balanced Multi-Class Dataset. University of Regina, Department of Computer Science. (2023).

[23] Mohamed A. Mahdi, SulimanMohamed Fati, MohamedA.G.Hazber, Shahanawaj Ahamad and Sawsan A.Saad: Enhancing Arabic Cyberbullying Detection with End-to-End Transformer Model. Computer Modeling in Engineering & Sciences,141(2), p1652-1671, (2024).

[24] R.Shrestha, R.Dave: Machine Learning for Identifying Harmful Online Behavior: A Cyberbullying Overview. Journal of Computer and Communications, 13(1), p26-40, (2025).

[25] H.Bouchal, A.BELAID: Arabic Hate speech and social networks offensive language detection . Information Processing at the Digital Age Journal. 27(2), p: 24-29, (2023).

[26] A.Mousa , I.Shahin , A.Bou Nassif , A.Elnagar: Detection of Arabic offensive language in social media using machine learning models. Intelligent Systems with Applications. 22 (200376), (2024).

[27] A.Faraj Alqahtaniand, M.Ilyas: A Machine Learning Ensemble Model For The Detection Of Cyberbullying. International Journal of Artificial Intelligence and Applications (IJAIA),15(1), (2024) .

[28] A.Root, L.Jakubowski, M.Vanamala: Exploration and evaluation of bias in Cyberbullying Detection with Machine Learning. arXiv preprint arXiv: 2412.00609. (2024).

[29] J.Wang, X.Xu, P.Yu, Z. Xu: Hierarchical Multi-Stage BERT Fusion Framework with Dual Attention for Enhanced Cyberbullying Detection in Social Media. arXiv:2503.00342. (2025).

[30] S.Sihab-Us-Sakib , R.Rahman, S.Alam Forhad , A.Aziz :Cyberbullying detection of resource constrained language from social media using transformer-based approach. Natural Language Processing Journal. 9(100104), (2024).

[31] A.Kumar, G.Sthanusubramoniani, D.Gupta, A.R.Nair, Y.A.Alotaibi,, M.Zakariah,: Multi-task detection of harmful content in code-mixed meme captions using large language models with zero-shot, few-shot, and fine-tuning approaches. Egyptian Informatics Journal, 30(100683). (2025).

[32] M.Azzeh, B.Alhijawi, A.Tabbaza, O.Alabboshi, N.Hamdan, D.Jaser: Arabic cyberbullying detection system using convolutional neural network and multi-head attention. International Journal of Speech Technology.Vol 27, p 521–537,(2024).

[33] Monirah Al-Ajlan, and Mourad Ykhlef : Firefly-CDDL: A Firefly-Based Algorithm for Cyberbullying Detection Based on Deep Learning. Computers,Materials & Continua.75(1) .(2023).

[34] S.Chen, J.Wang , K.He : Chinese Cyberbullying Detection Using XLNet and Deep Bi-LSTM Hybrid Model. Information,15(93). (2024).