Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(1), 26-40
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Machine Learning for Identifying Harmful Online Behavior: A Cyberbullying Overview

Rikin Shrestha, Rushit Dave

Computer Information Science, Minnesota State University, Mankato, USA
Email: rushit.dave@mnsu.edu

How to cite this paper: Shrestha, R. and Dave, R. (2025) Machine Learning for Identifying Harmful Online Behavior: A Cyberbullying Overview. *Journal of Computer and Communications*, **13**, 26-40.
https://doi.org/10.4236/jcc.2025.131003

Received: December 15, 2024
Accepted: January 19, 2025
Published: January 22, 2025

Abstract

In this modern era, platforms for digital/social media and video games are growing daily. People are becoming dependent on them from all ages and with many positive aspects, but there are drawbacks as well, one of which is cyberbullying. Cyberbullying is a form of bullying that uses technological platforms to bully others. It has effects on victims mentally, emotionally, and physically, which include low self-esteem, acting violently, despair, increased stress/anxiety, depression, self-harming/suicide, etc. Findings from this research study justify that it affects young people more, impacting their emotional development and overall safety. Real-time cyberbullying detection identifies and protects the target from further abuse and its effects. This study aids in determining the seriousness of the issue and the vulnerabilities that individuals can take advantage of to bully others. Additionally, it will help to understand how various features of cyberbullying detection function assist in developing a strong and trustworthy system and making a healthy online community. Natural Language Processing (NLP) models assess the textual content and analyze hashtags and comments. Similarly, image context is analyzed using Optical Character Recognition (OCR), which converts images into a machine-readable format for further examination. There are also Deep Neural Network models, such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BLSTM). CNN is utilized for text/picture classification, LSTM is used for long-term dependency learning, and BLSTM expands the network's input by encoding data in both forward and backward directions. Classifiers like Support Vector Machine (SVM) and Naïve Bayes help detect cyberbullying. A working cyberbullying detection system can detect cyberbullying on multiple platforms. A deeper understanding of each machine learning algorithm allows one to build a model that improves upon their predecessors. With models being developed for different attributes providing results with high accuracy, the cyberbullying detection system contributes by leading us to a healthier online community.

Keywords

Cyberbullying Detection, Natural Language Processing (NLP), Text Classification, Online Safety, Multilingual Detection Systems

1. Introduction

Internet use, particularly digital/social media and video games, is growing daily. Everyone is becoming dependent on them from all ages and with many positive aspects, there are drawbacks as well, one of which is cyberbullying. Cyberbullying has effects on victims mentally, emotionally, and physically. It includes low self-esteem, acting violently, despair, increased stress/anxiety, depression, self-harming/suicide, etc. Findings from this research study justify that it affects young people more, which impacts their emotional development and overall safety. With the increasing number of cyberbullying, researchers have developed various machine-learning models and detection systems to identify cyberbullying in real-time. These studies range from detecting bullying in text/image in social media, online gaming, and multimedia content.

Gaming platforms like DOTA and Ragnarok's data set were used to explore and research using the Convolutional Neural Network (CNN) model to detect cyberbullying, achieving 99.86% accuracy and showing high potential for detecting cyberbullying outside social media [1]. In addition to gaming, other research studies detect cyberbullying in multilingual form with a main focus on Hindi and Marathi, with the research taking in linguistic diversity and applying features such as Naïve Bayes and Logistic regression to get high accuracy [2].

As technology evolves so do speech patterns and behavior, changing the concept of bullying. Studies focusing on social media have shown that detecting abusive language alone cannot be considered bullying or be considered bullying. Using algorithms such as Support Vector Machines (SVM) and Logistic Regression, we can differentiate between actual cyberbullying, which can produce high accuracy in the detection of cyberbullying [3]. Gender-specific patterns of aggression also offer another option to explore, as research found that males and females express hostility differently online, improving detection accuracy and identifying the person being the bully [4].

With many detection models available for analyzing textual content, bullying detection systems have moved also to analyze multimodal content. The development of Optical Character Recognition (OCR) models allows for detecting cyberbullying in images and screenshots, broadening the scope of detection to non-text formats [5] [6]. Likewise, using stack embedding such as BERT has also provided positive results in new forms of text-based platforms to detect cyberbullying [7].

By combining textual and visual features, researchers have developed models capable of detecting cyberbullying across different platforms, even social media like Instagram [6]. These multimodal models offer a more comprehensive view

and understanding of negative online behavior, making them a gateway for the future development of cyberbullying detection systems. With the increasing number of social media platforms, this research outlines the contributions of machine learning to identify and prevent cyberbullying using various models for various forms. The various algorithms for analyzing text, images, and multimodal data help us understand the limitations of various models and the improvements required to create a robust model. This research focuses on the overview of detecting cyberbullying using machine learning, guiding a path to detect real-time bullying with high accuracy, and ensuring the safety of online communities.

2. Background

Cyberbullying can happen in many forms: harassment, racial comments, blackmail, offensive and vulgar language, spreading false rumors, and impersonation. With the rise of the internet and digital platforms, more and more people are becoming victims of cyberbullying. Just as a roof protects a house from rain, a cyberbullying detection system can prevent bullies from harming others.

Cause and effect: As technology increases so does its risk. The rise of new platforms comes with many benefits, but that benefits both good and bad, where people can misuse it to bully others. 46% of U.S. teens from age 13 to 17 have experienced cyberbullying at least once in a survey conducted on April 14-May 4, 2022 [8]. The Centers for Disease Control and Prevention (CDC) data shows that out of the adolescents that have been cyberbullied 13.6% have made a serious suicide attempt. Other individuals who were also victims had serious psychological impacts like increased levels of depression, anxiety, and the potential for long-term mental health issues. Research like this aims to shed light on the seriousness of this issue and help in developing a detection system to prevent further cyberbullying [8].

An active detection system will help prevent cyberbullying on a larger scale throughout multiple platforms like social media, online gaming, etc. Active models including Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and Natural Language Processing (NLP) can help in detecting and classifying large data with high accuracy [3]. With the platform allowing images, Optical Character Recognition (OCR) models can translate the image into machine-readable text for further processing, leaving fewer options for bullies to exploit [5]. Despite their advancements, we are still far from being able to prevent or detect cyberbullying completely. For example, distinguishing between actual bullying and sarcasm is challenging, as sarcasm can be flagged as bullying, resulting in false outcomes. Analyzing videos, other forms of multimedia content, or images with hidden meanings can also go undetected. Using multiple languages in a single sentence can hinder a model's ability to detect harmful content [2] properly. Achieving real-time detection remains another limitation in developing a robust model. To achieve the goal of being able to detect cyberbullying without flaws, these limitations need to be addressed.

The research helps prevent cyberbullying by developing machine learning models capable of detecting various methods used by bullies. By using large data sets from conversations and ensuring high accuracy in real-time detection, this research contributes to advancing better models, ensuring that victims receive the protection they need. It also provides insight into various machine learning algorithms and their pros and cons in detecting cyberbullying, while posing important questions: "What are the current limitations in detecting cyberbullying? How can we improve it?" and "How can these models be applied to detect real-time bullying, and what actions should be taken after it is detected?" Much research is being conducted on speech patterns and gender-specific language, which will help identify bullies using fake accounts to harass victims. This research also examines the different tactics bullies use, identifying other options that can be exploited to harm others.

3. Method

Building an effective cyberbullying detection system using machine learning involves several essential steps to ensure the highest possible accuracy. The method includes gathering data, preprocessing it, selecting the optimal machine learning algorithm, training, testing, and classifying results. These steps are critical for producing a functioning model capable of accurately detecting instances of cyberbullying.

Based on the process shown in **Figure 1** the data for the research papers was gathered mostly from Kaggle, where the datasets are both labeled and unlabeled. Depending on the dataset, two different types of algorithms are used for labeled data sets. Supervised Learning Algorithms are used, whereas Unsupervised Learning Algorithms are used for unlabeled data sets. The labeled datasets were primarily used because they provide clear and specific information and patterns that help the model learn distinctions. Labeled data also reduces uncertainty, increases model performance, is faster and easier to train, and simplifies the evaluation of the model's performance [9]. In contrast, an unlabeled dataset creates more challenges, as the model must first discover information and patterns without guidance. With this type of data, the results become less accurate and require additional steps to achieve comparable performance with models trained on labeled data. Not all datasets are from Kaggle; some research papers use data collected from game chat logs to detect cyberbullying [1]. Some research papers gather data from online gaming platforms like DOTA, Ragnarok, and World of Tanks, including diverse types, such as textual data from game chat logs [1]. Similarly, datasets were gathered from social media platforms such as Twitter, Instagram, and YouTube. The datasets from social media were in two forms: text or image [6] [9]. In some research papers, data was gathered in languages other than English [2]. The collected data was then manually labeled according to the study's needs; for instance, if the goal was simply to detect bullying, the labels used were binary ("0" and "1"). However, in some research papers, bullying was categorized in multiple forms, such as gender-based, blackmail, or racial, and was labeled accordingly before starting the modeling process [10].

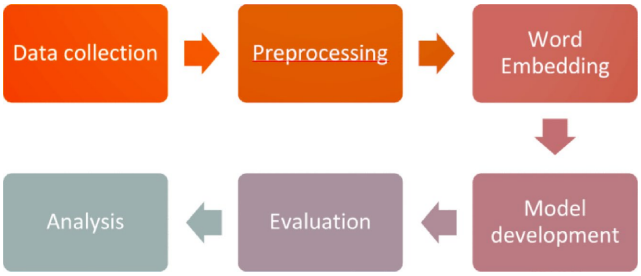

Figure 1. Steps for cyberbullying detection [1].

The datasets gathered will typically contain large amounts of data, and this data is often only labeled and raw, coming directly from the source. Depending on the source, the dataset may contain unnecessary pieces of information that need to be removed to improve the classification of the data. This is the first and crucial step that also helps to avoid unnecessary errors and processing in the model, allowing it to detect cyberbullying more accurately. The data gathered from game chat logs often contains extraneous information, such as in-game announcements like "user23421 has just accomplished the hidden task," which occurs frequently from start to finish and does not contribute to cyberbullying detection [1]. By preprocessing the data (**Figure 2**), we can remove these types of irrelevant information, reducing the model's load and making it faster. Similarly, data from social media may also contain unnecessary elements, like emojis, which can be removed [9]. Some research papers even go further by removing nouns, adjectives, pronouns, and common words like "a," "the," "has," "us," "if," "to," "in," "is," and "of," as these words are generally not effective for detecting cyberbullying. Removing them reduces data size and improves the model's speed and performance [3] [11].

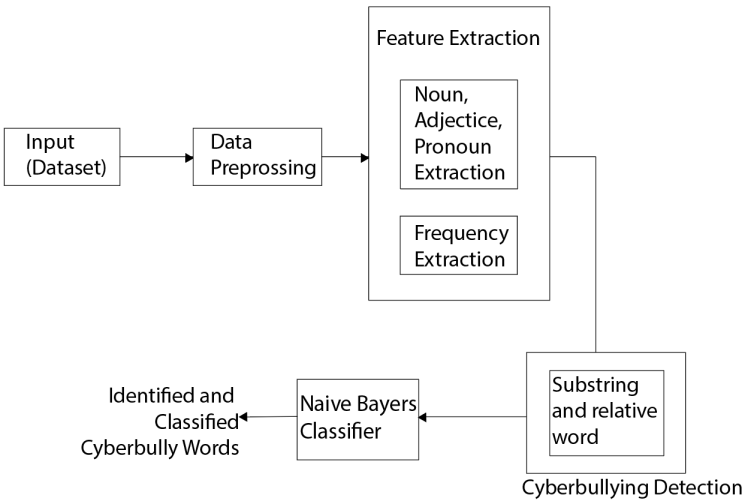

Figure 2. Model with feature extraction [11].

Natural Language Processing (NLP) is a frequently used field that can handle large text datasets. Not all data is textual; some may be image-based. However, NLP is not designed to read image-based data, so Optical Character Recognition (OCR)

(Figure 3) is used. This program scans and converts the image into text so NLP can process the data. This program is used mostly for multimedia analysis [9]. Classifiers such as Support Vector Machines (SVM) (Figure 4) are highly compatible with ORC, providing fast and efficient work in the detection of cyberbullying. The ORC is first applied to extract the text from the image then it is preprocessed for classification with the help of CNN. CNN identifies patterns to classify whether or not it is bullying. The classification by CNN can be binary ("1 and 0" or "bullying and non-bulling") or multiclass, where it's classified deeper on various forms of bullying like racial, gender-based, blackmail, etc. This can be used to detect bullying from an Instagram page that posts harmful posts detecting from both text and image and classifying it according to what type of bullying it is.

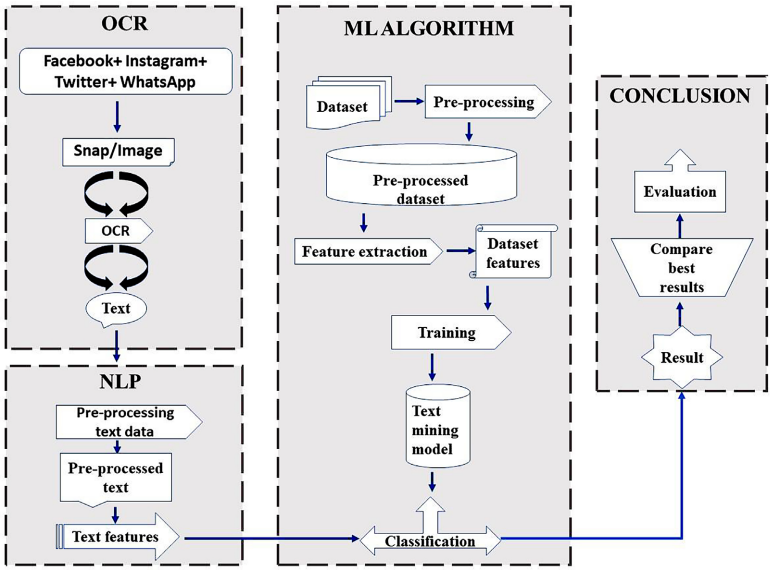

Figure 3. Model with OCR implemented [9].

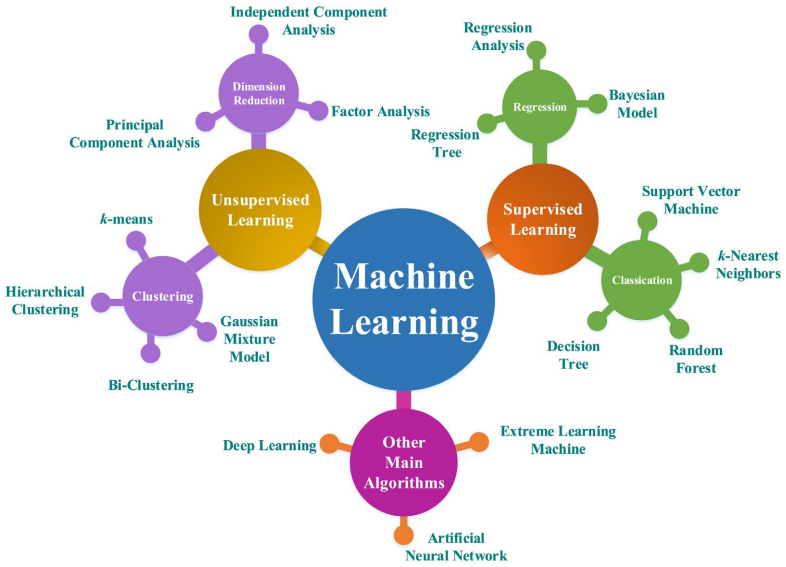

Figure 4. Machine learning classifier [12].

3.1. Supervised Learning Algorithms

Supervised Learning Algorithms (Figure 5) can only use labeled data to train the model and classify it as cyberbullying or not. This algorithm uses the labeled data as its reference to understand the pattern and then differentiate it between bullying and non-bullying.

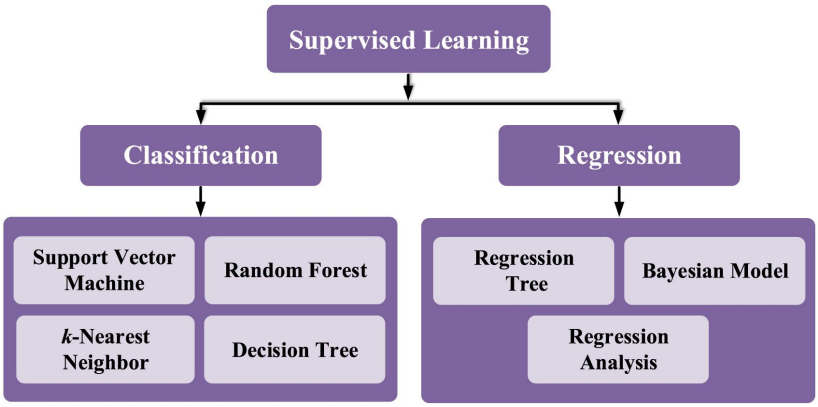

Figure 5. Supervised learning algorithms [12].

With the dataset pre-processed and labeled, we pick a model suitable for the data set then, we move on to the training and testing phases. The data is divided into two parts: one for training and one for testing, normally, it is split into an 80:20 or 70:30 ratio, where 80% or 70% of the data is used for training the model, leaving the rest of them untouched for testing. During the training phase, the model is taught to learn and recognize patterns associated with cyberbullying. Using a large dataset provides more nuanced patterns and improves the model's understanding. Some research papers utilize methods like Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BoW), and BERT to convert text content into numeric representations, enabling better pattern recognition in detecting bullying behavior [1] [13] [14].

For data classification, a binary approach is often used: 0 or 1, representing bullying and non-bullying. However, some research goes beyond binary classification, adopting a multi-class approach for more detailed categorization based on the model's requirements [1] [10]. One such classifier is the Convolutional Neural Network (CNN), it is highly compatible in image analysis with Optical Character Recognition (OCR) and sequence data, allowing it to recognize patterns efficiently [9] [14]. Among the most frequently used classifiers are Support Vector Machine (SVM) and Naïve Bayes. SVM, primarily used in binary classification, is highly effective with text-based data, demonstrating high accuracy in distinguishing bullying from non-bullying content [1]. For example, if we analyze Twitter comments first we preprocess the labeled dataset by removing unnecessary words like "a", "the", "we", etc. which is not useful when detecting cyberbullying. Then the SMV will be trained with the labeled data for a binary classification "bullying" as 1 and "non-bullying" as 0 making the model prepared to detect cyberbullying using

SMV. Naïve Bayes, on the other hand, is a probabilistic classifier based on Bayes' theorem: "P(A|B) = P(B|A) * P(A)/P(B)." This method calculates probabilities, enabling fast classification even on large datasets [10]. The dataset is vectorized using a method like TF-IDF, and then the dataset is divided into two sets: training and testing 80% and 20%, respectively. Other methods can also be implemented for Supervised Learning Algorithms such as decision trees, random forest, and logistic regression.

Decision tree: It is easy to understand and implement as it is good for classification and regression. It creates a tree structure where nodes in the tree represent the decision of whether it's cyberbullying or not. As the nodes increase, the tree and the last nodes are known as leaf nodes. It also uses both text-based and numeric data and predictions. Converting the text content into numeric representations can also be done. The drawback of using a decision tree is that it overfits; it creates an overly complex model that is not used for unused data [9] [15]. The implementation of the decision tree starts with converting the text data into numerical using BoW or TF-IDF. Then the tree construction is when data is split into branches forming a tree with two sides bullying and non-bullying. When it comes to classifying it traverses the tree based on the value to conclude.

Random forest: It is a learning algorithm that creates multiple trees unlike a single tree in a decision tree. It improves the classification, prevents a limitation of the design tree of overfitting, and creates a less complex model. It can also handle large data sets and high-dimension data, making it a great choice for dealing with text data. Due to the nature of random forests, the process requires multiple computations to function making it unsuitable for real-time detection [15]. Its function is similar to a decision tree, instead of a single tree, it contains multiple trees where it aggregates the prediction of all the trees to determine the final classification.

Logistic regression is an algorithm mostly used for binary classification. It predicts the probability of the classification, making it easy to implement and efficient. It has also proven to work well with a linear relationship. In contrast, it is limited when handling non-linear relationships [3]. The first step in implementing a logistic regression is to convert the text into a numeric feature using TF-IDF. Then the labeled data is split into two parts and used to train and test the model to fit a logistic regression model. Then testing phase, it computes the probability of the data and which class it belongs to and concludes the highest probability.

3.2. Unsupervised Learning Algorithms

Unsupervised learning algorithms (**Figure 6**) are used in the classification of unlabeled datasets. This method analyzes the data set to discover hidden patterns and clusters in the given data, then classifies them into bullying and non-bullying. Due to the nature of unsupervised learning algorithms using unlabeled data, it takes a longer time to process data and the accuracy is lower than supervised learning algorithms most papers prefer to use labeled data sets and the algorithm

associated with it [7]. Some papers use unlabeled data sets they gathered, in which they implemented algorithms such as K-Means Clustering.

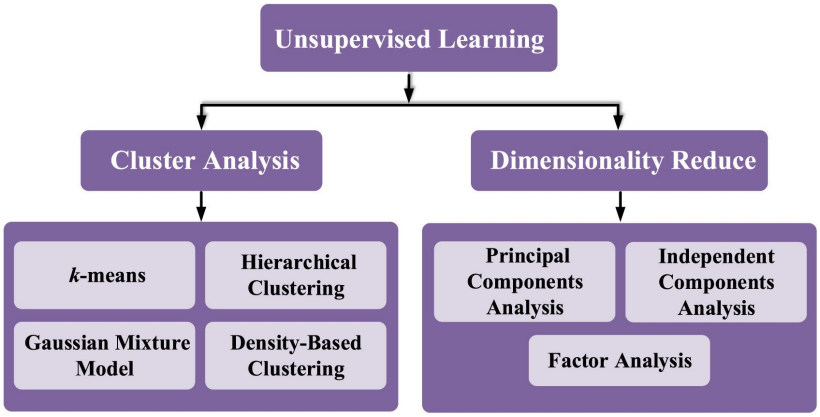

Figure 6. Unsupervised learning algorithms [12].

K-Means Clustering: This algorithm groups the dataset into different clusters based on a similar cluster. By taking the center of a cluster, it groups the data in k-distance grouping all the data clusters by classifying them into bullying and non-bullying. This method helps in handling large datasets and unlabeled datasets. Since the data are not labeled this method it's less accurate than algorithms that use labeled data sets. Since the data is not labeled this process will take longer, first the text data is converted into numeric vectors using TF-IFD. Then, the data is clustered using k-means to group it into clusters based on similarity. In the clusters cluster one can contain bullying and cluster two contains non-bullying. Those clusters were formed using the frequency of abusive words and the sentiment score of the data. Then, it analyzes the cluster to determine whether the new data is bullying or not.

4. Model Validation

Once the model completes training, the remaining 20 to 30% of unused data is used for the testing phase. A performance matrix is used to evaluate the model's effectiveness in classifying data, providing four key metrics: Accuracy, Precision, Recall, and F1 Score. Accuracy measures the percentage of correct classifications between bullying and non-bullying cases, offering an overall performance score [15]. Precision calculates the ratio of correctly identified bullying cases among true and false positives, giving a more nuanced insight into detection accuracy. On the other hand, recall indicates the ratio of true positives to the total of true positives and false negatives, where a higher recall suggests better detection capability for cyberbullying [15]. The F1 Score combines both Precision and Recall, incorporating all four factors—true positives, false positives, false negatives, and true negatives—often summarized in a confusion matrix [16].

True positive (TP): Correctly identified cyberbullying.

False positive (FP): Incorrectly identified bullying.

True Negative (TN): Correctly identified as non-bullying.

False Negative (FN): Missed bullying case by the model. [17]

$$\text{(Precision) } Pr = TP/(TP + FP)$$

$$\text{Recall} = TP/(TP + FN)$$

$$F1 = (2 * Pr * Pc)/(Pr + Pc) \text{ [17]}$$

In one of the papers [16] (**Figure 7**), they used a tweet from Twitter with a trace of bullying and applied it to their model. They used SVM and Naive Bayes, the accuracy was 71.25% and 52.70%, respectively. **Figure 6** represents the confusion matrix based on the result of our testing data [16].

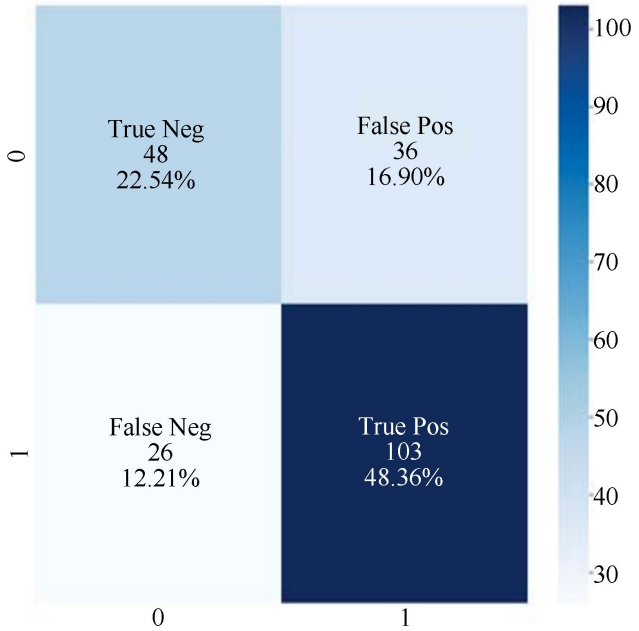

Figure 7. Confusion matrix [16].

5. Dataset

The given dataset (**Table 2**), based on variety of applied algorithms (**Table 1**), contains data from Twitter, Formspring, Facebook, etc., analyzing the size of the data and whether it is labeled or not. This dataset collection includes data in multiple languages like English, Hindi, Turkish, and Bengali, allowing multilingual functionality. These datasets are used to build various cyberbullying detection system models.

Table 1. Critical analysis on a different approach.

Approach	Analysis		
	Strength	Weakness	Best use for
Convolutional Neural Network (CNN)	Used for both textual and image-based data. Has a high accuracy.	It requires large data and needs to be well labeled. Required high training time.	Image/textual detection with a large dataset.

Continued

Support Vector Machines (SMV)	Effective for small datasets. Has high accuracy.	Scalability issues	Binary text classification.
Naïve Bayes	Fast and scalable, efficient for large data sets. Great for textual classification	Makes assumptions that might not align with the model.	Large textual dataset.
Random Forest	Used to handle large datasets. Reduces overfitting common setbacks in a single tree model.	It takes a long time to compute and does not apply to instant detection.	Large and high-dimension dataset.
Logistic Regression	Simple and efficient making it a great choice for binary classification.	Works with data with linear relationships.	Binary classification with linear patterns
K-Means Clustering	Can handle unlabeled data. Identifies grouping of data.	Low accuracy. Its cluster instability: the result may depend on the cluster	Unlabeled dataset.

Table 2. Dataset for a cyberbullying detection system.

Dataset Description/Link	Dataset properties		
	Dataset Type	Data Size	Source
https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset?select=toxicity_parsed_dataset.csv	Labeled	160,000 instances; size (66.72 MB)	Kaggle, Twitter, Wikipedia Talk pages, and YouTube
Ragnarök [1]	Labeled	534,328 lines	Ragnarök
Dota [1]	Labeled	230,394 lines	Dota
https://research.cs.wisc.edu/bullying/data.html	Labeled	534,950, released in June 2015	Twitter
https://github.com/eimearfoley/CyberBullyingDetection/blob/master/data/formspring.csv	Labeled	Size 3.79 MB	Formspring
https://github.com/karan19100/Cyber-Bullying-Detection-in-Hinglish-Languages-Using-Machine-Learning-/blob/main/Dataset/final_dataset_hinglish.csv	Labeled	Size: 4.69 MB	N/A

Continued

https://github.com/am-shb/cyberbullying-detection/blob/main/data/cyberbullying_tweets.csv	Labeled	Size: 6.84 MB	Twitter
https://www.kaggle.com/datasets/gbiamgaurav/cyberbullying-detection	Labeled	Size: 50.41 MB 115661 lines	Kaggle, Twitter, Wikipedia Talk pages, YouTube
https://github.com/snigdhab7/TextSecureAI/blob/master/negative.txt https://github.com/snigdhab7/TextSecureAI/blob/master/positive.txt	Labeled	Size: 3.14 KB-negative And 617 KB-Positive	N/A
https://www.kaggle.com/datasets/mrtbeyz/trke-sosyal-medya-paylam-veri-seti	Labeled	Size: 903.23 KB 11006 lines	Tweets in Turkish
https://www.kaggle.com/datasets/moshiurrahmanfaisal/bangl-cyber-bullying-dataset	Labeled	Size: 603.81 KB. 6010 lines	YouTube, Facebook, and Twitter in Bangali
https://www.kaggle.com/datasets/madhubalaji/cyberbullyingcsv	Labeled	Size: 7.17 MB 46017 Lines	Twitter
https://www.kaggle.com/datasets/mdkhurshidjahan01/cyberbullying-datset	Labeled	Size: 7.62 MB 54272 lines	Kaggle
https://www.kaggle.com/datasets/ishan8055/cyber-bullyinghinglish	Labeled	Size: 5.22 MB 17068 Lines	Kaggle

6. Limitation

Despite various models and algorithms providing high accuracy, cyberbullying detection systems still have many flaws and limitations. Just as life evolves, technology and human behavior take new steps in its evolution. With each passing generation, new words and speech patterns emerge. As people grow and change, so must our models. Detection systems must adapt to these changes in speech patterns to detect bullying in real time, which requires continuous training of the models. Similarly, as technology evolves, new online platforms emerge. Existing models may not be able to detect cyberbullying on these new platforms, requiring the creation of new models. This means that a single model cannot be universal—it is designed to work with the specific platform or dataset it was trained for.

Platforms such as TikTok and Snapchat focus primarily on images and videos, so models need improvement in decrypting photos, videos, and audio, and converting them into textual formats for further processing. Bullies may also use cryptic text in images with the intent to evade AI detection and continue harassing victims. With a model that can operate across multiple platforms and accurately decipher audio, video, and images, we can take the next step in detecting cyberbullying.

Another limitation of cyberbullying detection systems is distinguishing between sarcasm or jokes and actual bullying. Conversations between friends often include words or phrases that could be interpreted as bullying but are intended positively. For example, "I hate your fucking brain sometimes, but it is so useful" uses abusive language but is meant as a compliment [3]. One research paper used Support Vector Machines (SVM), a well-known and efficient binary classifier, to train their model. Logistic regression was employed to select the best combination of features. The SVM algorithm learns a classification function using training data. They selected some of the best content-based and sentiment-based features for their machine learning algorithms, as sentiment-based features consider the emotion of the text [3]. While this approach accounted for sarcasm and language complexity, the detection system's accuracy was not as high as that of other models.

There are multiple languages worldwide, and many people speak more than one language. Since cyberbullying detection systems are typically trained for a specific language, they often classify multilingual text as non-bullying. The use of aggressive words in different languages, combined with harmless English words, can bypass these detection systems [17] [18]. Multilingual detection systems have been developed to detect cyberbullying in languages such as Hindi, Marathi, Latin, and Arabic, but these systems still face limitations [7] [17]. Even with detection systems capable of processing multiple languages, the issue that comes with it is that they are often unable to handle emojis effectively [2]. With the use of aggressive words in different languages, people can use emojis by combining them into a harmful meaning like using black emojis for racial comments [18].

Even after solving these limitations, one of the main limitations comes down to the real-time detection of cyberbullying. The best and most effective prevent cyberbullying is to detect the content that contains bullying or now before the other person gets it. The creation of a robust detection system that can be implemented anywhere to prevent cyberbullying will make the online environment safer.

7. Conclusions

The continuous increase in cyberbullying with the increase in technological advancement and being more accessible to various age groups highlights the urgent need for a detection system to make the online environment safer. The effects of bullying on the victims cause major damage to them. This review addresses its problems, steps toward its solution, and the limitations of current methods to detect bullying. With AI taking the next step toward technological advancements, we will be able to create a system that will allow us to develop models to help detect cyberbullying. Machine learning models using algorithms such as Convolutional Neural Networks (CNN) and Natural Language Processing (NLP) have shown results with high accuracy in the detection of bullying, with classifiers such as Support Vector Machines (SVM) and Naïve Bayes. With models to detect textual

context, the limitation of bullying through images came up, which was countered using Optical Character Recognition (OCR), which extracts the text from an image into a machine-readable form.

Despite advanced models' progress and development, the detection system still faces many obstacles. The obstacles faced are differentiating between sarcasm and actual bullying and being able to detect it in multiple languages with a single model. This limitation underlines the reason for advancement in a more robust detection system that will lead to the development of real-time detection of bullying. Likewise, improving the accuracy of models with their ability to handle multimodal content by combining and more accurately extracting information from visual and audio data into machine readable form, helping take a step toward a comprehensive detection system.

Overall, this research is an overview of the pros and cons of various models with the potential and current limitations of cyberbullying detection systems. Looking at the restrictions, potentials, and benefits of the cyberbullying detection system, we can say that a robust real-time detection system will make cyberbullying a thing of history. By using what we have learned, we will be able to make better models and learn more from them and one day create a robust detection system that will make the online community safe and healthy.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cornel, J.A., Christian Pablo, C., Marzan, J.A., Julius Mercado, V., Fabito, B., Rodriguez, R., *et al.* (2019) Cyberbullying Detection for Online Games Chat Logs Using Deep Learning. 2019 *IEEE* 11*th International Conference on Humanoid*, *Nanotechnology*, *Information Technology*, *Communication and Control*, *Environment*, *and Management* (*HNICEM*), Laoag, 29 November-1 December 2019, 1-5. https://doi.org/10.1109/hnicem48295.2019.9072811

[2] Pawar, R. and Raje, R.R. (2019) Multilingual Cyberbullying Detection System. 2019 *IEEE International Conference on Electro Information Technology* (*EIT*), Brookings, 20-22 May 2019, 40-44. https://doi.org/10.1109/eit.2019.8833846

[3] Perera, A. and 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

[4] Dadvar, M., Trieschnigg, D., Ordelman, R. and de Jong, F. (2013) Improving Cyberbullying Detection with User Context. In: Serdyukov, P., *et al.*, Eds., *Advances in Information Retrieval*, Springer, 693-696. https://doi.org/10.1007/978-3-642-36973-5_62

[5] Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M.A., Yaseen, Q. and Gupta, B.B. (2024) Image Cyberbullying Detection and Recognition Using Transfer Deep Machine Learning. *International Journal of Cognitive Computing in Engineering*, **5**, 14-26. https://doi.org/10.1016/j.ijcce.2023.11.002

[6] Singh, V.K., Ghosh, S. and Jose, C. (2017) Toward Multimodal Cyberbullying Detection. *Proceedings of the* 2017 *CHI Conference Extended Abstracts on Human Factors*

in Computing Systems, Denver, 6-11 May 2017, 2090-2099. https://doi.org/10.1145/3027063.3053169

[7] Mahlangu, T. and Tu, C. (2019) Deep Learning Cyberbullying Detection Using Stacked Embbedings Approach. 2019 6*th International Conference on Soft Computing & Machine Intelligence* (*ISCMI*), Johannesburg, 19-20 November 2019, 45-49. https://doi.org/10.1109/iscmi47871.2019.9004292

[8] Atske, S. and Atske, S. (2024) Teens and Cyberbullying 2022. Pew Research Center. https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/#fn-92711-1

[9] Sultan, T., Jahan, N., Basak, R., Jony, M.S.A. and Nabil, R.H. (2023) Machine Learning in Cyberbullying Detection from Social-Media Image or Screenshot with Optical Character Recognition. *International Journal of Intelligent Systems and Applications*, **15**, 1-13. https://doi.org/10.5815/ijisa.2023.02.01

[10] Sifath, S., Islam, T., Erfan, M., Dey, S.K., Islam, M.M.U., Samsuddoha, M., *et al.* (2024) Recurrent Neural Network Based Multiclass Cyber Bullying Classification. *Natural Language Processing Journal*, **9**, Article ID: 100111. https://doi.org/10.1016/j.nlp.2024.100111

[11] Nandhini, B.S. and Sheeba, J.I. (2015) Cyberbullying Detection and Classification Using Information Retrieval Algorithm. *Proceedings of the* 2015 *International Conference on Advanced Research in Computer Science Engineering & Technology* (*ICARCSET* 2015), Unnao, 6-7 March 2015, 1-5. https://doi.org/10.1145/2743065.2743085

[12] Gao, K., Mei, G., Piccialli, F., Cuomo, S., Tu, J. and Huo, Z. (2020) Julia Language in Machine Learning: Algorithms, Applications, and Open Issues. *Computer Science Review*, **37**, Article 100254. https://doi.org/10.1016/j.cosrev.2020.100254

[13] Mahlangu, T., Tu, C. and Owolawi, P. (2018) A Review of Automated Detection Methods for Cyberbullying. 2018 *International Conference on Intelligent and Innovative Computing Applications* (*ICONIC*), Mon Tresor, 6-7 December 2018, 1-5. https://doi.org/10.1109/iconic.2018.8601278

[14] Hamiza Wan Ali, W.N., Mohd, M. and Fauzi, F. (2018) Cyberbullying Detection: An Overview. 2018 *Cyber Resilience Conference* (*CRC*), Putrajaya, 13-15 November 2018, 1-3. https://doi.org/10.1109/cr.2018.8626869

[15] Siddhartha, K., Kumar, K.R., Varma, K.J., Amogh, M. and Samson, M. (2022) Cyber Bullying Detection Using Machine Learning. 2022 2*nd Asian Conference on Innovation in Technology* (*ASIANCON*), Ravet, 26-28 August 2022, 1-4. https://doi.org/10.1109/asiancon55314.2022.9909201

[16] Desai, A., Kalaskar, S., Kumbhar, O. and Dhumal, R. (2021) Cyber Bullying Detection on Social Media Using Machine Learning. *ITM Web of Conferences*, **40**, Article ID: 03038. https://doi.org/10.1051/itmconf/20214003038

[17] Bayari, R. and Bensefia, A. (2021) Text Mining Techniques for Cyberbullying Detection: State of the Art. *Advances in Science, Technology and Engineering Systems Journal*, **6,** 783-790. https://doi.org/10.25046/aj060187

[18] Murnion, S., Buchanan, W.J., Smales, A. and Russell, G. (2018) Machine Learning and Semantic Analysis of In-Game Chat for Cyberbullying. *Computers & Security*, **76**, 197-213. https://doi.org/10.1016/j.cose.2018.02.016