Cyberbully Detection from Social Media with Text Classification

Zakaria Ibne Rafiq [24341150], Fahmid Hasan Chowdhury [21201286],
Rakibul Hasan Talukdar [23241100], Sabbir Hossain Mirza [23241086]
BRAC University Dhaka, Bangladesh

***Abstract*—With the rise of social media, cyberbullying has become a prevalent issue, causing significant emotional distress. Our model employs machine learning algorithms to categorize text based on demographic factors, including age, ethnicity, and religious beliefs. This review explores several papers that use machine learning to detect cyberbullying from social media like Twitter, WhatsApp, Youtube etc.**

***Index Terms*—Cyberbullying Detection, Twitter, Machine Learning (ML), Naive Bayes, Logistic Regression, Deep KNN, SVM, Natural Language Processing (NLP), Convolutional Neural Networks (CNN), Random Forest, Recurrent Neural Networks (RNN), Bidirectional GRU, Bidirectional LSTM etc.**

I. Introduction

The rise of social media has brought a wave of connection, instant news, educational resources, and entertainment. However, this digital landscape also fosters negativity, with cyberbullying being a particularly harmful issue. Cyberbullying is the use of electronic communication to bully a person, typically by sending messages of an intimidating or threatening nature. It can manifest in a multitude of ways, ranging from sending messages that are hurtful, mean, or threatening, spreading rumors or lies online, sharing photos or video footage embarrassing to the subject without permission, intentionally excluding someone from an online group or activity, and pretending to be someone with the intent to cause further distress. The impacts of cyberbullying on victims can be devastating and may include anxiety, depression, social isolation, and even self-harm.

Recent studies underline the scope of this growing problem. For instance, in a study conducted by the Cyberbullying Research Center found that 34% of students reported that they experienced cyberbullying in the past year.[1] Such harassment can severely affect a student's well-being and academic performance. Thus, there has been an ever-growing amount of research focused on applying machine learning algorithms and natural language processing techniques to detect cyberbullying. The purpose of this literature review is to review research that is related to the detection of cyberbullying detection systems, in particular, how these systems detect texts that are actually cyberbullying. One critical challenge people face is that the meaning of the word can vary depending upon its context. Certain words that can be often labeled indicators of cyberbullying, such as "filthy," "beast," and "wrecked," can also reflect positive emotions. Hence, the goal is to construct a model that is able to encompass those intentions..

This review will analyze various studies conducted to overcome the challenge of detecting cyberbullying from social media. By synthesizing findings from different approaches, this paper attempts to bring out best practices and various approaches used to help develop unbiased cyberbullying detection tools that contribute to a safer online environment. Ultimately, this review underscores the importance of data preprocessing and machine learning algorithms and tweaks that enhance the model performance, while creating a better understanding of the complexities involved in cyberbullying detection.

II. Literature Review

The literature review of the paper "Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms" [2] delves into some crucial topics surrounding the challenges and advancements in addressing cyberbullying online. First, the paper sets the stage by acknowledging the rise of aggressive behavior on social media as a direct consequence of digital communication technologies. Cyberbullying is a growing issue, and its rapid spread across online social networks (OSNs) presents complexities that traditional approaches cannot manage. The need for more scalable and effective solutions is urgent. The review highlights the importance of machine learning as a tool for predicting cyberbullying. Given the damaging effects of online aggression on mental, emotional, and social well-being, automating the detection and prevention process through predictive models is a key goal. Social media, with its vast reach, has become a breeding ground for bullying behaviors, and these platforms amplify their harmful impact. The paper acknowledges the difficulties in collecting relevant data from social platforms, especially when it comes to avoiding biases that stem from keyword-based searches. Getting a representative sample of social media interactions is crucial to developing effective machine learning models, but this task is easier said than done. Next, feature engineering is presented as a vital element in improving the performance of machine learning models. Features like age, gender, and the use of offensive language are often incorporated, but the dynamic and constantly evolving nature of language on social media makes this a challenging task. The review also compares various machine learning algorithms that have been used to detect cyberbullying, such as support vector machines (SVM), naive Bayes, random forests, and k-nearest neighbors. Each of these algorithms

has its pros and cons in terms of accuracy and scalability. SVM, for instance, often stands out for its high accuracy in classification. Moreover, evaluating these models is no easy feat, so metrics like accuracy, precision, recall, and the F-measure are used. Given the issue of class imbalance-where cyberbullying cases are much fewer compared to normal posts-the Area Under the Curve (AUC) is suggested as an additional evaluation metric to give a more balanced assessment of model performance. Lastly, the paper discusses ongoing challenges, such as keeping up with the ever-changing nature of human communication online. Slang, cultural differences, and the use of new acronyms create difficulties for building models that can universally predict cyberbullying. In essence, this literature review thoroughly explores the current state of research in predicting cyberbullying through machine learning. It identifies both the progress made and the obstacles ahead, particularly in terms of improving data quality, feature selection, and adapting to the fast-paced evolution of online behaviors.

The paper "Detection and Fine-Grained Classification of Cyberbullying Events" [3] explores the use of machine learning to automatically detect and categorize cyberbullying incidents using data from Ask.fm, comprising 85,485 Dutch-language posts. These posts were labeled to identify cyberbullying, the roles of participants (harasser, victim, or bystander), and specific types of bullying such as insults, threats, blackmail, defamation, curses, sexual content, defense, and support for the bully. The features extracted by the researchers included word unigrams, bigrams, character trigrams, and sentiment lexicons, all put through an SVM. Each post was treated as a binary classification task: cyberbullying or not. 10-fold cross-validation was used to ensure the accuracy of the results. When all features were combined, the F-score for cyberbullying detection reached 55.39%, with insults being predicted most correctly at 56.32%, while threats and defamation both resulted in very low F-scores: 19.84% and 7.41%, respectively, with encouragement toward the bully having the poorest showing due to the small number of posts. The highly imbalanced dataset, where most of the posts did not involve cyberbullying, made the study quite challenging. Besides, bag-of-words models could not capture subtle or implicit forms of bullying, like defamation and threats. Sentiment features also contributed little to the performance of the model. The authors believe that advanced linguistic features, such as syntactic and semantic patterns, would be of great use in the detection of implicit cases of bullying. They also suggest taking special care with issues such as data imbalance and increasing recall for crucial categories like threats and defamation. They also stress normalizing the informal and misspelled languages, which are habitual in social media posts for better model functionality.

The literature review of the paper "Analysis of Tweets for Cyberbullying Detection" [4] offers a detailed exploration of how machine learning (ML) and natural language processing (NLP) techniques are being used to tackle the growing problem of cyberbullying on social media platforms like Twitter. The authors, Mathur et al., propose a real-time detection system that integrates text analysis with image captioning, which stands out as a unique approach in this field. To start, the review highlights how machine learning has been employed for cyberbullying detection. Several previous studies have compared ML algorithms, such as Support Vector Machines (SVM), Naive Bayes, Decision Trees, and Random Forest, with varying degrees of success. For instance, Islam et al. found that while SVM is highly accurate, it can be too slow for real-time applications, a key issue this paper aims to address. Other researchers, like Ruminati et al., improved accuracy by including social context in their models, showing that cyberbullying detection benefits from more than just text data. The literature also touches on more complex models, like Zhang et al.'s Pronunciation-based Convolutional Neural Network (PCNN), which handled noisy data but struggled with precision due to unbalanced datasets. What sets this paper apart is its focus on multi-modal content detection, incorporating both text and images. The authors integrate image captioning techniques, using a neural model that evaluates whether the text accompanying an image matches the content of the image itself. This adds a new layer of analysis that many existing systems lack, especially when detecting harmful content in real-time. Additionally, with the real-time analysis, the authors emphasize the importance of timely detection. They use a tool called Selenium to scrape tweets and apply their machine learning model immediately, ensuring that new tweets are analyzed as soon as they are posted. In terms of algorithm performance, the paper compared several models, including Random Forest, AdaBoost, and Gradient Boosting, ultimately finding that an optimized Random Forest model worked best, achieving an impressive accuracy of 94.06%. The use of Count Vectorization and TF-IDF (Term Frequency-Inverse Document Frequency) techniques for feature extraction further enhanced the model's ability to capture the most relevant features from the data. Overall, this literature review acknowledges the progress made in cyberbullying detection but also points out the ongoing challenges, particularly in detecting more subtle forms of bullying like sarcasm or indirect insults. The inclusion of real-time analysis and multi-modal detection (combining text and images) in this paper pushes the field forward, offering a more holistic and practical solution for social media platforms to combat cyberbullying.

In paper [5] by Jason Wang et al. the authors wanted to create a machine learning model that is able to classify whether a cyberbully is targeting a victim's age, ethnicity, gender, religion, or other quality. One of the primary challenges in this domain is the severe class imbalance present in existing datasets, which can undermine the accuracy and fairness of classification models. In order to overcome that problem, the authors used a semi-supervised online Dynamic Query Expansion (DQE) process. This helped to create a more balanced dataset by extracting relevant data from Twitter. Additionally, the paper introduces a classification model based on a Graph Convolutional Network (GCN), which uses cosine similarity between tweet embeddings to construct a graph. The authors compiled their own dataset with 6 different twitter

based dataset that acted as the foundation which includes Agrawal, Bretschneider, Chatzakou, Davidson, Waseem, and WISC database. The dataset was further enhanced by manually classifying the tweets into fine-grained cyberbullying categories based on age, ethnicity, gender, religion, and other. They employed a semi-supervised online Dynamic Query Expansion (DQE) process to address the imbalance in the dataset and collect more data from twitter. They constructed a textual graph utilizing cosine similarity between tweet embeddings, forming a Semantic Cosine Similarity Graph Convolutional Network (SOSNet) to capture both local and global text data patterns. The authors preprocessed the data using various word embedding techniques—including Bag of Words, TF-IDF, word2vec, GloVe, fastText, BERT, DistilBERT, and Sentence BERT—and compared the performance of their model against several classifiers, such as Naive Bayes, SVM, XGBoost, Logistic Regression, k-Nearest Neighbors, and Multilayer Perceptron, using a 75:25 train-test split. After training a dataset of 40000 tweets shows a test accuracy of 94% and a F1 score of 94% resulted by XGB and BOW preprocessor. But on the other hand when only 4000 tweets are used the SOSNet model along with SBERT performed better with Test Accuracy upto 92% and a F1 score of 92%.

The paper "Cyberbullying Detection with Fairness Constraints" [6] tackles the challenge of detecting cyberbullying while reducing unintended biases in machine learning models. The researchers used four publicly available datasets: Jigsaw, focused on gender bias; a multilingual Twitter dataset, focused on language bias; Wikipedia Talk pages, centered on recency bias; and the Gab Hate Corpus, addressing religion, race, and nationality bias. These datasets were chosen to test the proposed fairness-constrained model in scenarios where cyberbullying and bias usually overlap. In the current study, a deep neural network was used with sentence-DistilBERT embeddings for binary classifications of cyberbullying. The key innovation herein was the embedding of fairness constraints at the training stage, which aimed at minimizing the bias against certain demographic groups. This is achieved by controlling FNED and FPED. These measures ensured that different groups were treated more equitably by the model. Results illustrated substantial bias reduction without sacrificing performance. Specifically, it reduces 78.7% of the total bias for the Jigsaw dataset, aimed at gender bias, while 94.9% for the Wikipedia dataset targeted at recency bias. On the other hand, Twitter data involved five languages and, therefore, showed only 11.3% bias reduction due to the complexity in handling out-of-vocabulary terms in multilingual datasets. The bias reduction is 27.5% for the Gab dataset, and most concerns hate speech on race and religion. Similarly, the models that underwent training with fairness constraints had higher results in MCC compared to baseline ones; Matthews Correlation Coefficient is a reliable metric when it comes to binary classification. As the number of constraints grows, so does the complexity of training and hence the difficulty of convergence. On the other hand, over-constraining can affect the performance of the model or simply lead to trivial solutions. Another limitation was the general unavailability of cyberbullying annotated datasets with demographic information, making the wider application of this model even further limited. Furthermore, imbalance within certain groups in the dataset, such as Twitter data domination with English posts, further barred the model from actually being multilingual. For future work, it remains to tune hyperparameters, look into more sophisticated neural architectures, and apply fairness constraints to other data modalities, such as images or videos. It would also have been very nice if there was some way of ascertaining the availability of more diverse and representative datasets and testing the results on a wide range of demographic groups.

The paper [7] addresses the development of an automated system for accurate detection of cyberbullying texts. They used a machine learning model called Bidirectional LSTM and used a feature selection algorithm called information gain. This approach tries to mitigate the challenge posed by the imbalance in the dataset of the cyberbullying classification. It also takes advantage of the contextual information within sentences in order to further improve accuracy. The paper focuses on how to overcome some of the key challenges that are related to noisy and imbalanced data, the common grounds in the social media text classification tasks. The paper is dedicated to discussing the impact of diverse preprocessing steps, including stopword removal and normalization, on the effectiveness of the model, trying to reveal that a careful selection of features may boost the detection performance. As for the dataset, they used an already collected dataset which is based on twitter provided by [5]. This paper uses the Information Gain (IG) algorithm for preprocessing data, focusing on identifying features that contribute to high accuracy in cyberbullying classification. IG calculates the amount of information a feature may contain, discarding irrelevant ones and retaining meaningful ones. It calculates the entropy of a class given a feature, reducing uncertainty in class predictions. However, this approach may lead to poor results when there is an imbalance in the dataset. To combat that issue they proposed an improved IG where they standardizes entropy into a range of $[0, \log_2 |c|]$, and factor inverts features concentrated in specific classes to prioritize them. Initial results showed that removing stopwords significantly improved precision, recall, and F1-score, with the best overall performance achieved using lemmatization resulted in Precision, Recall and F1-score of 94.54%, 94.50% and 94.52%. An information gain threshold of 0.0004 yielded optimal results with Precision, Recall and F1-score of 95.15%, 95.14% and 95.14% due to its ability to capture contextual semantics.

The authors of paper [8] stated that cyberbullying happens in all different contexts; hence, its category is hard to choose manually. So, they have combined multi-classification models in order to detect cyberbullying. They used the dataset which was created on Twitter before it rebranded as X and includes approximately 48,000 tweets from users of different ages, genders, cultures, and beliefs all over the world. The tweets were categorized according to age, gender, ethnicity, religion,

”other” for different kinds of cyberbullying, and ”not cyberbullying.” In their pre-processing, the authors used the NLTK library. They extracted n-grams with TF-IDF. They combined three classic classifiers: Decision Tree, Random Forest, and XGBoost, using ensemble methods like voting and stacking to increase accuracy. For the performance evaluation, accuracy, precision, recall, and AUC were used. In this study, five classifiers in total have been used. Among the multi-classifiers, the high of 89% was achieved by a Random Forest model using unigrams. On the other hand, the results for the ensemble methods of voting and stacking were even higher, at 90% and 90.71%, respectively. Overall, the accuracy of the models ranged from 86% to 90%. While the framework exhibited very impressive accuracy in cyberbullying detection, there were some shortcomings as regards capturing subtle word meanings. Also, because it is based exclusively on data from Twitter, there is a question as to its effectiveness on other media. Moreover, only English was used in this study; work on other languages is intended for future research. Another limitation is that it has only examined text and excluded cyberbullying expressed by emojis or images, a point the researchers say they will consider later. This paper concludes by proposing a novel framework that combines an ensemble of transformer models in detecting cyberbullying on Twitter and achieves an appreciable level of accuracy despite some limitations.

The researchers in paper [9] have not proposed any model themselves but have instead proposed a powerful system of Convolutional Neural Network for the detection of cyberbullying by overcoming the limitations of traditional methods. The paper points to the disturbing growth in cyberbullying on social media platforms like Facebook and Twitter, with particular emphasis on taking a serious toll on mental health thus demanding immediate automated detection. These systems use machine learning to detect and report cases of bullying, thereby protecting the victim indirectly. The authors also suggest that future researchers explore new avenues by exploiting the datasets containing cyberbullying texts from different social media sites. Regarding the methodology of the paper, the proposed work has adopted a CNN. In CNN, pre-processing of data has to be done very precisely. For this purpose, the authors have utilized NLTK that converts the input texts into sequences of word indices. It is composed of text categorization and converting into vector representations, while the architecture is a sequential CNN with an integration of activation functions and word embeddings. Though the model is compiled and trained and fit, accuracy with cyberbullying detection is determined later, hence providing a concrete skeleton for the implementation of the CNN system. Moreover, in this paper, no real-world scenario application and testing of this system are discussed or elaborated on, especially regarding questions on practical feasibility and scalability. What's more, online interactions could be considered in research to improve the effect of bullying detection. It is similarly important that such systems need constant updating in order to outsmart tactics changing every day by cyberbullies. This research positions the CNN system as a promising and versatile tool in fighting against cyberbullying, while highlighting a number of potential benefits for a range of stakeholders and emphasizing that further evolution is needed with regard to online harassment challenges.

The paper [10] proposes a new approach for the Gated Expansion Causal Convolution combined with Bidirectional Gated Recurrent Units for machine learning-based detection. Given the increasing concern for cyberbullying, the automatic approach aims to alleviate its psychological impact with better accuracy in the detection of harmful content. The experiments involved two datasets: a publicly available Twitter dataset from GitHub, consisting of 4,817 instances of cyberbullying and 6,037 examples not regarded as cyberbullying, and a new Chinese Weibo dataset prepared for this paper, containing 3,000 instances of each type, collected from comments. The architecture of the GECC-BiGRU model includes an input layer, a GECC layer for feature extraction, a BiGRU layer for sequence processing, an attention mechanism, and an output layer. Attention mechanism is critical; it underlines those features relevant for making more accurate predictions by paying attention to the most critical parts of the text. Regarding performance, among the best performances, the One-Layer model gave an accuracy and F1 score of 82.25% on the Weibo dataset, whereas in the Twitter dataset, the Three-Layer model was a bit better, with a score of 84.66% against 84.73% for the One-Layer model. However, this study has the major limitation of focusing solely on text-based detection and hence missing out on the important visual cues such as images and videos on social media. For future applicability, the model needs to be enhanced with multimodal analysis and expansion of the dataset with diverse and real-time content. Furthermore, refining the attention mechanism to adapt to evolving bullying language will help the model keep pace with the changing dynamics of online harassment. Overall, the paper demonstrates the effectiveness of the GECC-BiGRU approach in detecting cyberbullying text, outperforming existing methods while highlighting areas for further improvement.

The paper [11] focuses on implementing cyberbullying detection using BERT to create word embedding which are then fed into a single neural network layer used for classification of text. This paper highlights the power of BERT's transformer based architecture, especially the attenuation mechanism which was initially introduced by Google in 2017, to overcome the deficiencies of traditional models like RNN's. They choose to use BERT for its exceptional performance while retaining contextual and semantic comprehension of text. The authors used publicly available dataset from services like Youtube, Whatsapp, Twitter and had an expert annotate the combined dataset based on specific criteria. They collected a total combined data of 370000 comprising 2 classes which are cyberbullying and not cyberbullying. They used 93000 samples of data from each class to create a combined dataset of 186k instances. While preprocessing the data they used several popular preprocessing techniques. This includes the removal of foreign words and languages. This ensured that the data is purely based on the English language. Next they removed

instances that were no more than 3 to 4 words (too short). They normalized the data by removing numbers, punctuation, stop words, URL's , and extra white spaces that do not contribute to the classification. But they retained some information which includes tags, hashtags, links, emojis which they believed might provide important information regarding the intent of the message to classify. So, they used several tools to provide meaning to the emojis like a happy face emoji is replaced with "happy face". They used BERT which uses Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) during pretraining, and the model is enhanced with a dense layer which utilizes softmax activation for classification. This resulted into an Accuracy of 83%, Precision of 82%, Recall of 70% and F1 score of 76%.

The paper "Deep KNN Based Text Classification for Cyberbullying Tweet Detection" [12] proposes a machine learning approach to detect cyberbullying in tweets, integrating deep learning techniques with a k-nearest neighbors (k-NN) classifier for improved text classification accuracy. The study addresses the challenge of recognizing and preventing cyberbullying, which is exacerbated by the ubiquity of social media platforms like Twitter. Previous research has used techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and ensemble methods to detect cyberbullying in textual data. Researchers have applied various algorithms, such as CNNs trained on pre-trained word embeddings, to identify abusive language in vast text corpora like Twitter datasets. Studies show that such neural network-based models can achieve high accuracy, precision, and recall when applied to text classification tasks, but they require substantial computational resources. Furthermore, hybrid models for enhanced cyberbullying detection have also been explored, combining deep learning algorithms with other classifiers, demonstrating that hybrid approaches outperform individual models. However, the challenge remains in optimizing these models for real-time application and cross-language detection, particularly when dealing with languages other than English. Text classification methods have been central to cyberbullying detection, with word embeddings being a primary tool for representing text data. Recent advancements, such as stack embeddings that combine multiple embedding methods, have proven effective in cyberbullying detection tasks. While deep learning methods like CNNs and RNNs have shown promise in cyberbullying detection, traditional machine learning techniques (e.g., k-NN, support vector machines) remain competitive when combined with deep learning architectures. The hybrid approach of combining k-NN with deep learning aims to address some of the limitations of previous models. The proposed model uses deep layers of artificial neural networks (ANN) to enhance the feature extraction process, improving the accuracy of text classification tasks. By applying k-NN to the output of the deep learning model, the study seeks to exploit the advantages of both methods, achieving higher precision and recall in identifying cyberbullying tweets. The paper also highlights the efficiency gains of using k-NN in combination with deep learning. The training time for the k-NN model progresses at a rate of 0.001 seconds per minute, and the use of the Adam optimizer for model training further enhances convergence speed. Additionally, using the k-NN-based system reduces the time and computational resources required by traditional cyberbullying detection systems by 38% to 42%.

The paper "Effective Automatic Cyberbullying Detection Using a Hybrid Approach SVM And NLP" [13] presents a hybrid method for detecting cyberbullying by combining machine learning and natural language processing (NLP) techniques. This research is based on the HateSpeech and Offensive Content Identification in Twitter Dataset (HOCTID), which includes over 140,000 tweets labeled as either cyberbullying or non-cyberbullying. The approach tends to use NLP for text preprocessing and tokenization, leveraging linguistic and psychological insights from the texts. Some of the features utilized in this research to enhance detection include Term-Frequency Inverse Document Frequency (TF-IDF) and the Linguistic Inquiry and Word Count (LIWC2) tool. TF-IDF identifies key terms while LIWC2 extracts emotional and psychological indicators from the text, which help the Support Vector Machine (SVM) classifier make a difference between cyberbullying and non-cyberbullying posts. The SVM model presents very promising results with an accuracy of 93.15%, beating all other models such as Random Forest, RNN, and CNN. It also scores high precision and recall scores at 0.93 and 0.96, respectively. Nevertheless, the limitations that are mentioned in the paper concern the imbalanced datasets and the complex linguistic pattern which can sometimes lead to misclassifications. The authors go on to say that feature extraction should be refined, along with the consideration of deep learning models, which may be able to capture the subtlety of languages that traditional methods fail to provide. They also mentioned two other promising areas of research that include tackling imbalance in data and enhancing real-time detection capability.

III. Conclusion

The literature review of cyberbullying detection using machine learning identifies great strides toward this pressing concern of social networking sites. The studies have pointed out the importance of scalable, real-time detection systems to reduce the adverse impacts of cyberbullying. Some of these different approaches have utilized machine learning models including Support Vector Machines, Random Forests, and Convolutional Neural Networks while considering issues such as class imbalance, feature selection, and the dynamic nature of online language. With all these developments, there is still room for improvement in adapting to the changing online behaviors, ensuring model predictions are nondiscriminatory, and improving subtle forms of bullying detection. Future research will need to be directed more toward enhancing the multimodal analysis, real-time application, and cross-platform generalization to make those models robust and effective in diverse social contexts.

REFERENCES

[1] S. Hinduja, "Cyberbullying statistics 2021 — age, gender, sexual orientation, and race," 2021, accessed on: 2024-04-27. [Online]. Available: https://cyberbullying.org/cyberbullying-statistics-age-gender-sexual-orientation-race

[2] M. A. Al-Garadi *et al.*, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges," *IEEE Access*, vol. 7, pp. 70 701–70 718, 2019.

[3] C. V. Hee *et al.*, "Detection and fine-grained classification of cyberbullying events," in *Recent Advances in Natural Language Processing*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:4245513

[4] S. A. Mathur *et al.*, "Analysis of tweets for cyberbullying detection," in *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 2023, pp. 269–274.

[5] J. Wang, K. Fu, and C.-T. Lu, "Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 1699–1708.

[6] O. Gencoglu, "Cyberbullying detection with fairness constraints," *IEEE Internet Computing*, vol. 25, no. 1, pp. 20–29, 2021.

[7] M. Xin, J. Shen, and P. Hao, "Cyberbullying detection and classification with improved ig and bilstm," in *2022 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, 2022, pp. 259–262.

[8] A. F. Alqahtani and M. Ilyas, "An ensemble-based multi-classification machine learning classifiers approach to detect multiple classes of cyberbullying," *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 156–170, 2024. [Online]. Available: https://www.mdpi.com/2504-4990/6/1/9

[9] S. M. Kargutkar and V. Chitre, "A study of cyberbullying detection using machine learning techniques," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 734–739.

[10] X. Li, H. Zhou, and C. Cui, "Cyberbullying detection based on gecc-bigru," in *2023 International Conference on Intelligent Management and Software Engineering (IMSE)*, 2023, pp. 169–172.

[11] R. Sujud *et al.*, "Cyberbullying detection using bidirectional encoder representations from transformers (bert)," in *2024 IEEE International Mediterranean Conference on Communications and Networking (MedItCom)*, 2024, pp. 257–262.

[12] M. Nisha and J. Jebathangam, "Deep knn based text classification for cyberbullying tweet detection," in *2022 11th International Conference on System Modeling Advancement in Research Trends (SMART)*, 2022, pp. 1550–1554.

[13] J. Sathya and F. M. Harin Fernandez, "Effective automatic cyberbullying detection using a hybrid approach svm and nlp," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 2024, pp. 1–6.