# Detecting Cyber Bullying: A Review of Techniques and Applications

Parom Guha Neogi, Tahsin Zaman Jilan, Rafsan Zamil and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
parom.guha.neogi@g.bracu.ac.bd, tahsin.zaman.jilan@g.bracu.ac.bd, rafsan.zamil@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—This review study provides an in-depth overview of both historical and contemporary research on recognizing cyberbullying. The study looks at several methods, including keyword-based strategies, machine learning methods, and social network analysis, for detecting cyberbullying. The dynamic and erratic character of online communication and the difficulty in accurately identifying and diagnosing cyberbullying behavior are two issues mentioned in the research with regard to detecting cyberbullying. The report also makes recommendations for additional research, such as the requirement for bigger datasets and the development of uniform evaluation metrics. Overall, this book highlights the significance of continued innovation in this crucial area and provides insightful information about the current level of research into the detection of cyberbullying.**

***Index Terms*—Cyberbullying, research, data, detection**

## I. Introduction

As more people use internet platforms to harass, threaten, and harm others, cyberbullying has become a serious concern. Finding cyberbullying examples is a difficult process that involves the use of modern methodologies and technologies to carefully locate and evaluate the underlying patterns and behaviors. This is especially true given how difficult it may be to construct good detection algorithms in the face of the dynamic and ever-changing nature of Internet communication. In this environment, the study of how to recognize and respond to occurrences of cyberbullying has become an important subject of study. Researchers and practitioners in this discipline are working to create cutting-edge approaches and technology.

Cyberbullying detection is a challenging and intricate subject with a number of approaches and methodologies. The most often used methods are keyword analysis, machine learning techniques, and social network analysis. Keyword-based detection is the process of looking for terms or phrases that are commonly used in cyberbullying, such as insults, threats, or abusive comments. Even with a wide range of tools, it might be difficult to identify cases of cyberbullying. This is mostly due to the fluidity and unpredictability of online communication as well as the difficulty in precisely identifying and diagnosing cyberbullying activities. The prevalence of anonymous or pseudonymous internet identities can also make it harder to track down offenders and hold them responsible for their actions. To better comprehend the intricacies and complexity of online communication, several researchers are investigating the use of natural language processing tools. Others are creating systems and tools to use social networks and crowdsourcing to find and report cases of cyberbullying. In general, the study of cyberbullying detection is essential to comprehending and combating this expanding social issue. Researchers and practitioners may contribute to the creation of safer and more respectful online environments for everyone by carrying out ongoing studies and the development of novel methods. This introduction provides an overview of the current state of research in this field, highlighting key challenges, approaches, and recommendations for future research.

## II. Cyberbullying effects and repercussions

Cyberbullying is a concerning phenomenon that has been brought on by the expanding usage of electronic communication tools. Using email, text messages, and social networking sites to maliciously damage another person is a particularly heinous form of harassment. Cyberbullying can have a variety of detrimental effects on both the victim and the offender. Cyberbullying victims could go through a lot of emotional and psychological pain. Regularly hearing insults and teasing can make a person feel anxious, hopeless, and even suicidal. Additionally, victims could struggle to establish and maintain relationships with others and may feel socially isolated. Additionally, cyberbullying may seriously harm a victim's academic performance, resulting in lower grades, more absences, and even dropping out. Both the victim and the abuser may have severe repercussions as a result of cyberbullying, including:

Emotional and psychological impact: Cyberbullying can drastically alter a person's life because of its horrible emotional and psychological effects. The constant flood of offensive and abusive letters that those who experience this type of harassment receive can destroy their sense of worth, confidence, and self-esteem. The effects could be severe and protracted, resulting in great emotional anguish and critical mental health problems. Because of their increased stress and strain, victims of cyberbullying may feel alone and helpless. They could begin to avoid once-loved activities, retreat from social interactions, and find it difficult to concentrate on daily duties. Cyberbullying can be especially damaging to people who already have mental health concerns. Victims may experience depression or an increase of the symptoms of their mental disorder. The negative repercussions of cyberbullying can sometimes be so severe that victims consider or attempt suicide. Cyberbullying produces emotional and psychological harm, which must be acknowledged as genuine and serious as physical harm. Those who have been victimized by cyberbullying require aid and resources in order to survive and heal. We must all work together to raise awareness about the hazards of cyberbullying and take actual efforts to prevent it in our communities.

Social isolation: Social isolation is among the most severe effects of cyberbullying. Online bullied and mistreated victims may avoid social situations and stop interacting with others. They might think they don't belong, are imperfect, or aren't deserving. The impacts of social isolation can be profound and pervasive. It may be difficult for victims to make and keep friends, which can result in feelings of isolation and loneliness. They might also experience a loss of social confidence, which would make it harder for them to speak up for themselves or engage in group activities. Lack of social connections can cause both physical and mental health to deteriorate over time, resulting in depressive, anxious, and low self-esteem feelings.

Legal consequences: Cyberbullying has long-lasting negative impacts beyond the victim's emotional and social wellness, and offenders may face serious legal repercussions. Legislators have recognized cyberbullying as a serious problem with detrimental effects. The perpetrator may be charged with both civil and criminal offenses, depending on how severe the bullying was. Cyberbullying can occasionally be classified as a hate crime, which carries harsh punishments. Civil lawsuits may also result in hefty fines and a requirement that the offender compensates the victim for any losses they sustained. A criminal conviction may also be followed by a prison term, probation, or other legal penalties. Cyberbullying can have legal ramifications outside of the courtroom. Cyberbullying can impair a perpetrator's reputation and employment prospects, making it more difficult for them to find work, maintain work contacts, and even pursue higher education. In today's digital age, decisions taken online can have significant and long-term implications.

Reputation damage: Cyberbullying can harm a victim's online and offline reputation in the long run. Cyberbullying can be extremely damaging to a person's reputation, hurting both their personal and professional lives in a variety of ways. It can permanently harm a victim's reputation by causing a loss of trust, respect, and social standing. Because of internet anonymity, bullies can abuse their victims without fear of repercussions. They may spread inaccurate, damaging, or embarrassing material about the victim on social media platforms, blogs, or websites, causing the victim's reputation and credibility to suffer. The proliferation of these harmful posts can be difficult to control, making repair challenging. In the business world, reputation harm may be highly costly. Online searches for information on a person may be conducted by prospective employers, coworkers, or clients. If they come across damaging or unattractive content about the victim, it can significantly influence their job prospects. Reputation harm can also have an impact on interpersonal relationships, making it more difficult to retain or create new bonds with previous acquaintances. Reputation harm can have long-term and, in some situations, irreparable implications. People may become victims, making it harder for them to move on with their life. It may be difficult to repair the damage done to their reputation because it will be costly.

The effects of cyberbullying on a victim's reputation outside of the virtual world can be severe and long-lasting. By creating a loss of trust, respect, and social status, cyberbullying can harm a person's personal and professional life. It is crucial that we understand the seriousness of the issue and take steps to prevent it. We can create a caring and compassionate online environment by encouraging responsible digital citizenship and developing an inclusive society.

## III. Literature review

The papers [1] inquiry is focused on the escalating issue of cyberbullying in virtual communities. The text expounds upon the deficiencies inherent in the current automated detection of abusive language and proposes an innovative framework known as Q-Bully. Integrating reinforcement learning and natural language processing techniques within this framework enhances the detection process.The report commences by drawing attention to the prevalent incidence of cyberbullying on digital platforms and its detrimental impact on the psychological and physiological well-being of individuals. The above statement underscores the crucial necessity for a viable and expandable resolution to address the issue of cyberbullying, given the infeasibility of manually overseeing vast quantities of information. The authors have acknowledged that the ever-changing nature of internet platforms poses a challenge to existing detection techniques. The literature review section introduces three key concepts, namely reinforcement learning, Q-learning, and natural language processing (NLP). Reinforcement learning is a machine learning methodology that centers on the attainment of predetermined goals. The methodology entails dispensing incentives or penalties to an algorithm predicated on its actions and decisions. The application of reinforcement learning in detecting occurrences of cyberbullying can be likened to the pedagogical approach of utilizing rewards or imposing punitive measures to promote a child's language acquisition abilities. The Q-Bully framework is founded upon the Q-learning algorithm, which is a reinforcement learning algorithm that is specifically targeted toward a particular objective. The methodology involves an agent utilizing a Q-table that incorporates the rewards associated with diverse actions across multiple stages for decision-making purposes. In order to attain a specific objective, the agent acquires the ability to optimize its decision-making mechanism for the purpose of selecting actions. The aforementioned segment pertains to the subject matter of natural language processing and its function in converting unstructured linguistic data into a format that is comprehensible to machines. The present context highlights two fundamental techniques in natural language processing, namely stop word elimination and stemming. Stemming and stop-word elimination are two techniques employed in the domain of natural language processing to enhance the analysis and processing of textual data. Stemming and stop word removal are two linguistic methods utilized in natural language processing. Stemming involves breaking down words into their fundamental forms, while stop-word removal entails the elimination of frequently used words that do not contribute to the semantic analysis or classification of textual data. This report provides a thorough description of the dataset utilized in the study, consisting of 184,354 comments that were meticulously categorized into two discrete groups: offensive and non-offensive. The authors make reference to the incorporation of additional baseline models and datasets to facilitate comparative analysis. The present study outlines the techniques employed for data cleansing on the dataset, encompassing word segmentation, special character elimination, stop word removal, and emoji handling. The manuscript provides a thorough presentation of the Q-Bully methodology, which employs reinforcement learning techniques to differentiate between comments that are deemed offensive and those that are not. The predictive model's operational mechanism entails treating each word in a sentence as a discrete state to ascertain the likelihood of its categorization as an offensive term. The Q-table is utilized to document the incentives associated with said actions. The algorithm's adaptability is emphasized by its capacity to handle dynamic vocabularies and changing contextual subtleties of lexicons. The utilization of the exploitation coefficient is proposed by the authors as a means to improve the convergence rate of the Q-Bully algorithm. The aforementioned coefficient enables precise state identification through the utilization of established lexical contexts. The process of creating a hash table entails the utilization of both benign and aggressive terminologies that are frequently utilized, while also factoring in analogous conditions by means of implementing stemming and Jaro-Winkler distance.

The scholarly article [2] details a collaborative endeavor that centers on malware analysis and showcases the most extensive collection of annotated malware reports that is accessible to the public on a global scale. This study emphasizes the growing significance of cybersecurity in the contemporary digital era and the potential efficacy of natural language processing (NLP) methodologies in contributing significantly to this domain. The authors engage in a scholarly discourse concerning significant cyber attacks, including the ”WannaCry” ransomware attack of 2017 and the Mirai botnet attack, highlighting the wide spectrum of malware threats. The insufficiency of annotated data in the domain of cybersecurity is recognized as a challenge for scholars in the field of natural language processing (NLP) who have a keen interest in this domain. The authors have presented a suggested approach for annotation methodology that aims to tackle the aforementioned challenge. The methodology involves the process of recognizing and labeling lexical units and phrases in malware reports that pertain to the operational features and functions of the malevolent software. The present discourse expounds upon the delineation of the three phases of annotation, namely token labeling, relation labeling, and attribute labeling. The dataset is known as MalwareTextDBv2.0 has undergone annotation and consists of 85 reports pertaining to Advanced Persistent Threat (APT), with a cumulative total of 12,918 sentences that have been annotated. The article outlines four distinct subtasks that are employed for the purpose of evaluation. The tasks involved in this study encompass the categorization of pertinent sentences, the anticipation of token designations, the anticipation of relation designations, and the anticipation of attribute designations for texts that are relevant to malware. Machine learning algorithms are utilized to provide baseline models for each subtask. The article concludes by providing a synopsis of the statistical data of the dataset, the metrics used for evaluation, and the models used as a baseline. The significance of collaborative endeavors in propelling natural language processing (NLP) research for cybersecurity is underscored by the authors. Additionally, they offer valuable perspectives on the difficulties faced throughout the annotation procedure. In summary, the manuscript presented herein makes a significant contribution to the advancement of natural language processing methodologies for the purpose of malware analysis. As such, it represents a valuable resource for scholars and practitioners engaged in research within the domain of cybersecurity.

This study [3] suggests an innovative neural network model for identifying online harassment in web content, to contrast the performance of deep neural networks with conventional machine learning algorithms, and to investigate the effects of various ways to extract features on the models' precision. The authors tested the proposed framework on two real-world cyberbullying datasets by assessing the effectiveness of the models, they developed a unique neural network architecture with parameter tuning and an algorithm comparison study of eleven categorization algorithms. The suggested method was evaluated using two real-world harassment datasets, and its value was determined by contrasting it to seven feature collection approaches using multiple classification strategies. Conventional machine learning techniques including Logistic Regression, Random Forest, Support Vector Machines, and Naive Bayes are employed in the paper's classification models, along with deep neural networks such as Convolutional Neural Networks, Recurrent Neural Networks, and Attention Models. The researchers discovered that while Logistic Regression was the best of the conventional machine learning models implemented, bidirectional neural networks and attention models also produced excellent categorization results with accuracy and F1-scores as high as 95% and 98%.

The paper [4] looks into the matter of cyberbullying on digital platforms, enhances current methods for identifying it, and creates a system with includes statistics visualization, ways to identify cyberbullying, as well as autonomous reporting, the study article uses machine learning algorithms including word2vec, LSTM, and CNN. To extract local characteristics and categorize the tweets as cyberbullying or not, the authors created an LSTM-CNN framework utilizing word2vec to train unique word embeddings. Additionally, they evaluated how well their strategy performed in comparison to other machine learning methods including Random Forest, Logistic Regression, and XGBoost. Finally, they created a website and Telegram chatbot that can determine whether or not a tweet constitutes bullying according to the degree of harm and help stop it. Except for XGBoost when it comes to ROC AUC, the LSTM-CNN model performs better than the other models across all measures with an accuracy of 95.2%.This suggests that the LSTM-CNN model outperforms the other algorithms in detecting abuse tweets.

This paper [5] aims to offer a unique way for spotting cyberbullying on social media platforms by implementing a deep learning model BERT. For the specific job, a pre-trained BERT model is employed with only one linear neural network level placed on top of the BERT model as a classifier which has been trained on the particular dataset in order to get the dataset-specific embeddings, and this model is evaluated using two internet-based datasets, one modest and the other a little larger in size. The paper compares the proposed approach to past studies that used deep learning models and traditional machine learning models alongside different word-embedded methodologies. This model's 12 layers of transformers are used to create the final embeddings. The given input data is only encoded in each layer using transformer encoders. The results show that the proposed technique outperforms past studies that used deep learning and conventional machine learning models with different word embedding strategies. The validation loss measure was used to keep the model that was trained from undergoing overfit, and several hyperparameters were used to evaluate the model's performance. Twitter datasets using CNN has an accuracy of 93.97% and the Wikipedia dataset has 96% accuracy in the BERT model.

This paper [6] aims to investigate the issue of cyberbullying and develop an independent linguistic model for text classification of cyberbullying. The authors highlight the spike in cyberbullying that has taken place since the COVID-19 epidemic and its negative effects on victims, including decreased self-worth and increased suicidal ideation. In earlier works, the identification and classification of cyberbullying were accomplished using the deep neural network (DNN) method known as Bidirectional Encoder Representations for Transformers (BERT). The section on related work discusses a variety of investigations into the identification of cyberbullying while emphasizing its limitations and platform-specific nature. The majority of research employs social context-based word representation techniques like Word2Vec, GloVe, and FastText.However, only a small number of OSNs, including ASK.fm, Twitter, Instagram, and Vine, are relevant to these investigations. Traditional machine learning techniques, such as Support Vector Machines (SVM) and Bi-directional Long Short-Term Memory (Bi-LSTM), have been used in several studies to represent language. For training and evaluation, datasets from a variety of OSN platforms, including Instagram, Vine, ASK.fm, and Formspring. me, and Twitter, are employed. The datasets were incorporated into the development of training, validation, and test sets. Using random oversampling approaches, the discrepancy between bullying and non-bullying episodes in the datasets is rectified. Linguistic models such as Bi-LSTM, HateBERT (a retrained BERT model), and SVM with TF-IDF were employed in the study's trials. Based on F1 ratings for both the positive (bullying) and negative (non-bullying) classes, the models are changed.

The article [7] discusses the expanding problem of cyberbullying, particularly among teenagers, and the need for technology to effectively recognize and prevent it. The authors point out the limitations of current research that evaluate all texts from all users uniformly without distinguishing between bullies and victims. The two objectives they emphasize in their novel methodology are 1) recognizing player roles as a multi-class classification issue and 2) identifying cyberbullying as a binary classification problem. The related works section lists earlier studies that looked at participant role identification and the classification of cyberbullying. Some studies have used user-based and social network-based features, while other research has focused on the lexical and semantic elements of participants' posts. By combining supervised learning techniques with already trained language models, the authors hope to reduce the requirement for rule-based approaches and task-specific feature extraction methods. This study adds to past role identification research. The two activities are covered in great detail in the model description section. For the classification of cyberbullying, the authors develop an ensemble model based on DistilBERT that was trained using a Twitter dataset for the detection of profanity. Three independent classifiers that were trained on unbalanced datasets of offensive and non-offensive tweets make up the ensemble model. Using a voting method, the predictions from the skewed classifiers are combined. The authors employ a BERT-based model with a pre-trained BERT embedding layer, a hidden neural layer, and a softmax output layer for role classification. The model is trained to separate postings into distinct roles before classifying posts into bullying and defending categories. The dataset utilized, the AMiCA dataset obtained from the social networking website ASKfm, is described in the techniques section. The dataset includes posts that have roles like ”harasser,” ”victim,” ”bystander defender,” and ”bystander assistant” tagged on them. In order to preserve identical data distribution ratios among the classes, the authors use stratified sampling and 10-fold cross-validation on the dataset.

The present investigation [8] is focused on the development of a web-based platform aimed at identifying instances of offensive language in Arabic. The present study endeavors to analyze the import of offensive language on the internet and its ramifications on the wider society. This statement underscores the exigency for additional inquiry in this domain. The current investigation delineates the involvement of the authors in the SemEval 2020 collaborative undertaking pertaining to Multilingual Offensive Language Identification. The present study involved an evaluation of the models employed by the authors, with a view to enhancing their performance. Additionally, an error analysis was conducted in order to identify areas for improvement. The present study endeavors to make a valuable contribution to the domain of multilingual offensive language identification. The present study was designed to investigate various deep learning models, namely RNN, GRU, Bi-GRU, LSTM, and Bi-LSTM. The objective of this inquiry was to ascertain the maximum attainable macro-F1 score. The findings of the present investigation demonstrate that the SalamNET system attained the maximum macro-F1 score of 0.83. The SalamNET system employed a Bidirectional Gated Recurrent Unit (Bi-GRU) model in tandem with Term Frequency-Inverse Document Frequency (TF-IDF) features. The present study offers a thorough examination of the existing literature on the detection of offensive language in Arabic. This study comprehensively examines various aspects of a research project, including the employed methodology, feature engineering techniques utilizing TF-IDF and AraVec word embeddings, baseline models, deep learning models, results, and error analysis. This paper endeavors to furnish a comprehensive survey of the extant literature pertaining to the detection of offensive language in Arabic, with the objective of identifying the lacunae in the present research. The present study's results hold potential implications for guiding future research endeavors in the field and enhancing the precision of Arabic offensive language detection models.

This study [9] aimed to perform binary classification of cyberbullying through the utilization of various classifiers and data augmentation techniques. The present study's results suggest that the uncased classifier based on Bidirectional Encoder Representations from Transformers (BERT) demonstrated superior performance in the context of the non-augmented dataset. In the context of the extended datasets, it was observed that the CNN classifier exhibited superior performance compared to other classifiers. The results of the study indicate that Logistic Regression models exhibited superior accuracy and F1 scores in comparison to Naive Bayes models. It is worth noting that the aforementioned models were surpassed in performance by the BERT and CNN models. The present study's findings suggest that the 'Character Level TF-IDF' feature outperformed the 'Word Level TF-IDF' feature in the context of TF-IDF analysis. The present study evaluated the efficacy of various data expansion techniques and found that the performance of the proposed methods was comparable. The present study found that Method 1 demonstrated superior performance compared to other methods. This outcome can be attributed to the method's capacity to identify and integrate sense-specific synonyms. The present study reports on the significant improvement in classification results observed for the extended datasets. The observed enhancement is ascribed to the efficacy of semantic meaning expansion methodologies, which were implemented via the utilization of disambiguation and Wordnet. This study presents a newly developed data augmentation technique that has demonstrated superior performance compared to the commonly utilized Mixup method. Empirical evaluations were conducted on various test sets, and the results consistently demonstrated the superior performance of the proposed technique. The present study investigated the efficacy of utilizing contextual meaning and synonyms as compared to random word replacements for the purpose of data expansion. The findings of the study revealed that the former approach resulted in superior outcomes. The present study conducted a comparative analysis with a prior investigation, which revealed that convolutional neural network (CNN) models trained on augmented datasets exhibited superior performance. The results of the classification process for the extended datasets exhibited a marked improvement in comparison to the initial datasets. The present study reports on the performance of a Convolutional Neural Network (CNN) on two distinct datasets, namely AskFm and FormSpring. The accuracy score for the CNN was evaluated on both datasets, and the results indicate a significant improvement in performance. Specifically, the accuracy score for the CNN increased from 91% to 94.3% for the AskFm dataset and from 95% to 98.3% for the FormSpring dataset. These findings suggest that the CNN model is effective in accurately classifying data from both AskFm and FormSpring datasets. The present investigation reveals that the suggested approach for data augmentation demonstrated enhanced efficacy in contrast to the Mixup methodology. This is supported by a 2.4% rise in the accuracy score.

The study [10] uses transferable learning to solve the issue of cyberbullying identification. The researchers employed a variety of small BERT models to fine-tune the simulations and included the Focal Loss function in order to tackle the disparities in the data. The researchers intended to provide cutting-edge findings in hate-speech identification by establishing that real-time applications of cyberbullying detection may be used with smaller BERT models. The authors evaluated the efficiency of their plan using 10-fold cross-validation on the complete dataset. The outcomes show that their approach is capable of outperforming prior work on the same dataset, despite without accounting for user- and network-based information. The smaller BERT models, which are suitable for real-time applications, were also shown to be faster in detecting harassment. The researchers achieved state-of-the-art performance on the hate speech dataset with 0.91 accuracy, 0.92 recall, and 0.91 F1 score.

This paper [11] investigates the employment of a supervised machine learning algorithm as the best method for identifying cyberbullying on Twitter-based networks because it has grown to be a big problem in the world of online networking, making its identification and prevention of the utmost importance. User behavior and tweet content must be entered into models to find trends in Twitter data, however, the development of such models is impeded by the lack of trustworthy and comprehensive datasets. The detection technique developed by the authors, which is based on decision trees, logistic regression, and support vector machines among other machine learning techniques, yielded good outcomes. To have a full sample set that illustrates both cyberbullying and non-cyberbullying behavior, the dataset must be balanced. They discover by testing several feature sets that the most effective combination is user behavior and tweet content. The examination of the detection system, which obtained an f-measure of 0.936 and an area under the receiver-operating characteristic curve of 0.943, demonstrates that it is very accurate in identifying cyberbullying tweets.

In this paper [12] researchers are looking into techniques of detection and prevention for this phenomenon as cyberbullying on social media has received increasing attention recently. An innovative solution to this problem is proposed in the research Automatic Detection of Cyberbullying by using a deep learning-based approach that is capable of not only identifying but also differentiating between various types of cyberbullies. It is challenging to recognize cyberbullying due to the diversity of content. Two examples of cyberbullying, which primarily consists of harassment, are threats and insults. The uneven difference between positive and negative data presented another challenge for the analysts. The researchers overcome these challenges by extracting temporal and contextual information from posts on several social media networks using a deep learning strategy that blends LSTM and CNN. The researchers overcome these challenges by employing the approach, although the dataset's imbalance between the number of positive and negative samples may have an impact on the accuracy of the findings. The proposed approach outperformed multiple sophisticated approaches, with an accuracy rate of up to 87%. Additionally, by utilizing CNN and LSTM networks as part of its operation, this model is capable of identifying pertinent components from input text that directly relate to temporal or contextual information included in postings. Without this crucial knowledge, it is difficult to identify and categorize the many forms of cyberbullying. Given the model's high levels of accuracy, deep learning-based methods to stop cyberbullying on social media seem like a promising choice.

This research paper [13] is an illustration of how effectively a hybrid deep learning technique can detect cyberbullying on social media platforms can be found in the paper Hybrid Deep Learning technique for Cyberbullying Detection on Social Media. A safer online environment might be achieved by using the suggested method, which has great accuracy in recognizing tweets featuring cyberbullying. The results of this study could influence the development of more reliable and efficient cyberbullying detection techniques on various social networking websites. In this study, a method called the Hybrid Deep Learning Strategy for Cyberbullying Detection on Social Media is offered as a remedy for the issues with machine learning-based cyberbullying detection. In order to categorize cyberbullying content, the authors' proposed hybrid technique involved developing a highly accurate cyberbullying detection system utilizing deep learning and natural language processing technology. To extract spatial features and time information from input texts, use the CNN and Balsam networks, respectively. The complexity of social media data and reliance on trustworthy feature extraction are two major issues covered in the research regarding developing a cyberbullying detection system. Additionally, the ability to extract both spatial and temporal data from social media posts is made possible by integrating the advantages of the CNN and Balsam networks in a hybrid approach. According to the study, by using hybrid-deep-learning-based algorithms, we can detect instances of cyberbullying on Twitter with an accuracy rate of up to 90% and a precision level of about 89%. Additionally, it is anticipated that the F-score and specificity scores will be between 88% and 91%, respectively. These results demonstrate that the proposed strategy outperforms the vast majority of existing alternative strategies for identifying cases of online bullying.

In this paper [14], a new method called Bully Net—which consists of three phases—is introduced for effectively detecting cyberbullying on Twitter. A signed network (SN) created especially for cyberbullying is also created, allowing us to examine bullying tendencies and examine tweets to determine their correlation with cyberbullying. When detecting those who engage in cyberbullying via an online social network, the optimized bullying score based on context when sending tweets outperforms other measures already in use. The system is evaluated using a dataset of 5.6 million tweets, and the results reveal that it is quite accurate at identifying cyberbullies while being scalable in terms of tweet volume. The authors discovered that when conversations are framed around context as well as topic, it is easier to properly identify the ideas and behaviors underlying bullying. The authors examined their proposed centrality metrics in their experimental investigation to isolate bullies from the signed network and discovered that they could do so with 70% accuracy and 77% precision. Finally, the BullyNet algorithm gives a practical method for identifying cyberbullies on Twitter, potentially aiding in the fight to reduce the negative effects of cyberbullying on social media platforms.

## IV. Findings

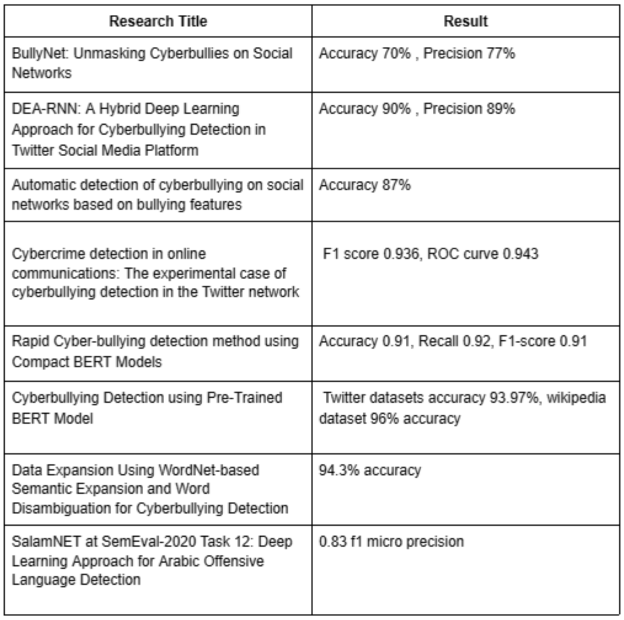
The important conclusions of this investigation were discovered following a thorough analysis and comparison of various scientific works. After carefully reviewing those publications, we discovered that authors in the field often employ a set of four critical methods. Every text classification approach uses these critical techniques: dataset collecting, data preprocessing, data partitioning, and feature selection.

Data collecting: Data collecting must be done with caution because cyber bullying is such a sensitive subject. Data collection on this subject may be difficult due to the topic's extreme sensitivity and sentimentality. To acquire a thorough picture of the presence and effects of cyberbullying, it is essential to explore a number of resources, including social media websites and online discussion boards. Nevertheless, it is critical to safeguard individuals' security and privacy when collecting data. To prevent unauthorized access or data breaches, efforts should be made to anonymize and de-identify any personally identifying information that may be contained in the data. As a result, acquiring information on cyberbullying needs a sophisticated and cautious technique that strikes a balance between the desire for comprehensive information and the commitment to protecting everyone's safety and privacy.

Data pre-processing: Although pre-processing data presents its own set of challenges, it is a necessary step in the interpretation of cyberbullying data. The large amounts of unstructured data produced by social media and other digital channels can be difficult to manage and correctly analyze, necessitating the employment of specialized methodologies and technologies. Ethical concerns must also be addressed in order to safeguard everyone's security and privacy. To address these problems, researchers may use pre-processing techniques including feature selection, data normalization, and data purification. These strategies aid in standardizing data formats, finding and removing superfluous or meaningless data, and extracting critical information for further investigation. By rapidly pre-processing data, researchers can gain a better understanding of the occurrence and effects of cyberbullying while simultaneously ensuring that ethical considerations are taken into account. Pre-processing data is a critical phase in the analysis of cyberbullying data that necessitates a determined plan and certain methods.

Partitioning data: Data partitioning is a crucial step in the study of cyberbullying data, but because the material is so delicate, there are some unique challenges. Researchers must constantly take care to preserve people's safety and privacy in order to assure the study's validity and accuracy. Utilizing stratified sampling to divide the data into proportionally represented sub-groups is one efficient method. Data on cyberbullying should be divided into training, validation, and testing sets for the most effective investigation. by taking the necessary precautions to secure individuals' privacy and safety while carefully examining the unique problems highlighted by this type of data.

Feature selection: The feature selection stage of the data analysis technique is crucial. Researchers must carefully identify the most significant features in order to distinguish instances of cyberbullying and explain the underlying patterns and behaviors. Using previously done cyberbullying research to identify relevant qualities, such as the usage of specific terms or phrases, is useful. Researchers can also employ sophisticated data analysis methodologies, such as machine learning algorithms, to uncover key features automatically. The ever-changing nature of social media and digital platforms complicates the selection of criteria for cyberbullying research. Researchers must constantly update and improve their feature selection techniques to stay up with changes in the internet world.



| Research Title | Result |
|---|---|
| BullyNet: Unmasking Cyberbullies on Social Networks | Accuracy 70% , Precision 77% |
| DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform | Accuracy 90% , Precision 89% |
| Automatic detection of cyberbullying on social networks based on bullying features | Accuracy 87% |
| Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network | F1 score 0.936, ROC curve 0.943 |
| Rapid Cyber-bullying detection method using Compact BERT Models | Accuracy 0.91, Recall 0.92, F1-score 0.91 |
| Cyberbullying Detection using Pre-Trained BERT Model | Twitter datasets accuracy 93.97%, wikipedia dataset 96% accuracy |
| Data Expansion Using WordNet-based Semantic Expansion and Word Disambiguation for Cyberbullying Detection | 94.3% accuracy |
| SalamNET at SemEval-2020 Task 12: Deep Learning Approach for Arabic Offensive Language Detection | 0.83 f1 micro precision |

Fig. 1. Reviewed papers results

## V. Conclusion

The study of cyberbullying has grown in popularity as internet harassment has intensified. Its goal is to safeguard vulnerable people from the negative impacts of cyberbullying by detecting and preventing such behavior. Because of developments in natural language processing and machine learning, new ways for precisely identifying and terminating cyberbullying have been developed. However, there remains a big barrier because cyberbullies are continually changing their techniques to avoid detection as online communication evolves. Despite this, scientists are constantly enhancing their algorithms to boost the accuracy and effectiveness of cyberbullying detection. We must keep up our efforts to combat cyberbullying and make the internet a better place for everyone. We can build on this achievement and eventually change the internet into a place where everyone can be more compassionate and kind by collaborating, exchanging information and resources, and continuing to grow.

## References

[1] A. T. Aind, A. Ramnaney, and D. Sethia, "Q-bully: A reinforcement learning based cyberbullying detection framework," in *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1–6.

[2] P. Phandi, A. Silva, and W. Lu, "SemEval-2018 task 8: Semantic extraction from CybersecUrity REports using natural language processing (SecureNLP)," in *Proceedings of the 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 697–706. [Online]. Available: https://aclanthology.org/S18-1113

[3] 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*, vol. 10, no. 22, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/22/2810

[4] M. Gada, K. Damania, and S. Sankhe, "Cyberbullying detection using lstm-cnn architecture and its applications," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 2021, pp. 1–6.

[5] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained bert model," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 1096–1100.

[6] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "HateCheck: Functional tests for hate speech detection models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 41–58. [Online]. Available: https://aclanthology.org/2021.acl-long.4

[7] G. Rathnayake, T. Atapattu, M. Herath, G. Zhang, and K. Falkner, "Enhancing the identification of cyberbullying through participant roles," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, Nov. 2020, pp. 89–94. [Online]. Available: https://aclanthology.org/2020.alw-1.11

[8] F. Husain, J. Lee, S. Henry, and O. Uzuner, "Salamnet at semeval-2020 task12: Deep learning approach for arabic offensive language detection," *arXiv preprint arXiv:2007.13974*, 2020.

[9] M. S. Jahan, D. R. Beddiar, M. Oussalah, and M. Mohamed, "Data expansion using wordnet-based semantic expansion and word disambiguation for cyberbullying detection," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 1761–1770.

[10] M. Behzadi, I. G. Harris, and A. Derakhshan, "Rapid cyber-bullying detection method using compact bert models," in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, 2021, pp. 199–202.

[11] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563216303788

[12] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proceedings of the 17th International Conference on Distributed Computing and Networking*, ser. ICDCN '16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: https://doi.org/10.1145/2833312.2849567

[13] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, and H. D. E. Al-Ariki, "Dea-rnn: A hybrid deep learning approach for cyberbullying detection in twitter social media platform," *IEEE Access*, vol. 10, pp. 25 857–25 871, 2022.

[14] A. S. Srinath, H. Johnson, G. G. Dagher, and M. Long, "Bullynet: Unmasking cyberbullies on social networks," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 2, pp. 332–344, 2021.