Detecting Cyber Bullying: A Review of Techniques and Applications

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 behaviour 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, Machine learning methods, Social network analysis, Detection algorithms, Natural language processing, Crowdsourcing, Safer online environments, Threats**

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:

A. 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.

B. 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.

C. 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.

D. 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 research by Aind et al. 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 [1]. The detection of cyberbullying and the reduction of its negative consequences are improved by the merging of reinforcement learning and natural language processing. The authors recognise the challenge of switching internet platforms. The literature review introduces reinforcement learning, Q-learning, and natural language processing (NLP). Similar to how kids learn to talk, reinforcement learning involves rewarding or punishing actions. The Q-Bully framework optimises decision-making through Q-learning. Stop word elimination and stemming are two NLP strategies that enhance textual data processing. In the report, the dataset, data cleaning procedures, and Q-Bully methodology are all explained. The authors suggest using the exploitation coefficient to increase convergence by taking lexical contexts into consideration. When building a hash table, stemming and Jaro-Winkler distance are utilised to precisely identify the terms.

The study by Raj et al. 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 investigating the effects of various ways to extract features on the models' precision [2]. In order to validate the proposed framework on two real-world cyberbullying datasets, the authors created a customised neural network architecture with parameter tuning and conducted an algorithm comparison analysis of eleven classification algorithms. The proposed method's performance was studied by comparing it to seven feature collection approaches that used various classification methods and were tested using two real-world harassment datasets. Convolutional neural networks, recurrent neural networks, and attention models were used in the study, which are deep neural networks combined with more traditional machine learning methods like logistic regression, random forest, support vector machines, and naive bayes.The researchers found that while Logistic Regression was the most effective of the common machine learning models used, attention models and bidirectional neural networks also produced well-categorized data with accuracy and F1-scores as high as 95

Gada et al. investigate cyberbullying on digital platforms, improves current methods for detecting it, and develops a system with statistics display, methods for detecting cyberbullying, and autonomous reporting [3]. The research paper makes use of machine learning techniques such as CNN, LSTM, and word2vec. The authors used word2vec to build an LSTM-CNN framework to train specific word embeddings, extract local characteristics, and categorize whether or not the tweets featured cyberbullying. 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.

Another study by Yadav et al. aims to offer a unique way of spotting cyberbullying on social media platforms by implementing a deep learning model BERT [4]. 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.

The study by Rottger et al. aims to investigate the issue of cyberbullying and develop an independent linguistic model for text classification of cyberbullying [5]. 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 over-sampling 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 research by Rathnayake et al. discusses the expanding problem of cyberbullying, particularly among teenagers, and the need for technology to effectively recognize and prevent it [6]. By viewing player roles as a multi-class classification problem and categorizing cyberbullying as a binary classification challenge, the authors propose a novel method for identifying cyberbullying. By merging supervised learning techniques with previously trained language models, they eliminate the necessity for task-specific feature extraction methods. They use a BERT-based model to categorize roles and an ensemble model based on DistilBERT to detect cyberbullying. Training is carried out using stratified sampling and 10-fold cross-validation on the AMiCA dataset from ASKfm.

The present investigation by Husain et al. is focused on the development of a web-based platform aimed at identifying instances of offensive language in Arabic [7]. This study highlights the need for additional research by studying the negative impacts of profanity on the internet and in general culture. The authors evaluate the models employed in the SemEval 2020 project on Multilingual Offensive Language Identification in order to enhance their performance and conduct error analysis. Deep learning models (RNN, GRU, Bi-GRU, LSTM, Bi-LSTM) are analysed to achieve the maximum macro-F1 score. The SalamNET system, which uses TF-IDF and Bi-GRU features, has a macro-F1 score of 0.83. In-depth analyses of methodology, feature engineering, models, results, and error analysis are provided for the corpus of research on Arabic offensive language detection in the report. The findings are important for developing Arabic offensive language detection systems and for next research.

The study by Jahan et al. aimed to perform binary classification of cyberbullying through the utilization of various classifiers and data augmentation techniques [8]. The CNN classifier outperformed other classifiers on extended datasets, but the BERT classifier excelled on non-augmented datasets, according to the study. Despite the fact that BERT and CNN models outperformed both, naive Bayes proved to be more accurate than logistic regression models. In comparison, word-level TF-IDF outperformed character-level TF-IDF. The semantic meaning expansion enabled by disambiguation and Wordnet dramatically improved categorization results. In terms of data growth, contextual meaning and synonyms surpassed random word swaps. On the AskFm and FormSpring datasets, the CNN model improved its accuracy, with scores increasing to 94.3 and 98.3%, respectively. The proposed data augmentation solution outperformed the mixup methodology by 2.4% in accuracy.

The study by Behzadi et al. uses transferable learning to solve the issue of cyberbullying identification [9]. 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.

Al-garadi et al. research study is the most effective method to identify cyberbullying on Twitter-based networks using a supervised machine learning algorithm because it has grown to be a major worry in the world of online networking [10]. User behavior and tweet content must be entered into models in order to find trends in Twitter data; yet, the creation of such models is hampered by the absence of reliable and comprehensive datasets. Positive results were obtained using the scientists' detection method, which is based on decision trees, logistic regression, and support vector machines, among other machine learning methods. 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 the research by Zhao et al. they were looking into techniques of detection and prevention for this phenomenon as cyberbullying on social media has received increasing attention recently [11]. 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 cyber bullies. 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.

In another research by Murshed et al. shows 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 [12]. 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 the study by Srinath et al. aims to introduce a new method called Bully Net—which consists of three phases—is introduced for effectively detecting cyberbullying on Twitter [13]. 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 pre-processing, data partitioning, and feature selection.

A. 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.

B. 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. Researchers may use pre-processing techniques including feature selection, data normalization, and data purification to address these problems. 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.

C. 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.

D. 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 behaviours. 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.

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

TABLE I
RESULTS FROM REVIEWED PAPER

Research Title	Result
BullyNet: Unmasking Cyberbullies on Social Networks [13]	Accuracy 70% , Precision 77%
DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform [12]	Accuracy 90% , Precision 89%
Automatic detection of cyberbullying on social networks based on bullying features [11]	Accuracy 87%
Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network [10]	F1 score 0.936, ROC curve 0.943
Rapid Cyber-bullying detection method using Compact BERT Models [9]	Accuracy 0.91, Recall 0.92, F1-score 0.91
Cyberbullying Detection using Pre-Trained BERT Model [4]	Twitter datasets accuracy 93.97%, wikipedia dataset 96% accuracy
Data Expansion Using WordNet-based Semantic Expansion and Word Disambiguation for Cyberbullying Detection [8]	94.3% accuracy
SalamNET at SemEval-2020 Task 12: Deep Learning Approach for Arabic Offensive Language Detection [7]	0.83 f1 micro precision

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] 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

[3] 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.

[4] 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.

[5] 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

[6] 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

[7] 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.

[8] 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.

[9] 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.

[10] 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

[11] 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

[12] 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.

[13] 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.