WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

The Study of Methods for Detection and Prevention of Cyberbullying

Prof. Manjunath Patil [a], ***Rohith Kawatagi*** [b], ***Rasale Shreya*** [c], ***Shirly Zakkam*** [d], ***Vinay Kadapure*** [e]

[a] *Department of Computer Science and Engineering, Angadi Institute of Technology and Management, Belagavi-590009, India*
[b] *Department of Computer Science and Engineering, Angadi Institute of Technology and Management, Belagavi-590009, India*
[c] *Department of Computer Science and Engineering, Angadi Institute of Technology and Management, Belagavi-590009, India*
[d] *Department of Computer Science and Engineering, Angadi Institute of Technology and Management, Belagavi-590009, India*
[e] *Department of Computer Science and Engineering, Angadi Institute of Technology and Management, Belagavi-590009, India*

A B S T R A C T

Social media users and their activity are both drastically increasing in today's world. These internet settings offer great chances for contact and knowledge transfer. Cyberbullying is a situation where some individuals abuse these tools to harass and abuse others online. It is crucial to identify cyberbullying as soon as possible to prevent any harm from being done to victims because of its negative impacts on people, especially young people (like PTSD-Post Traumatic Stress Disorder, anxiety, depression etc). To address this issue, a number of strategies and methods for identifying and preventing cyberbullying on social media have been created. In this study, we carried out a survey to look into the state of the art for cyberbully detection tools. Reviewing pertinent publications that have been published in scholarly journals and conference proceedings is a part of our study. According to the findings of our survey, numerous procedures and strategies are now being developed, and some of them are even being applied in specific locations. To assess their performance and identify the best solutions for various situations, more study is required because there is a lack of standardisation and comparability among these methods.

Keywords: Cyberbullying, Natural Language Processing, Frameworks, PTSD

1. Introduction

Social media is a term for computer-based technology that makes it possible to share information, ideas, and thoughts with others through online communities and networks. Internet based social media allows for instantaneous electronic sharing of content, including documents, movies, and images as well as personal information. Through web-based software applications, users interact with social media on a computer, tablet, or smartphones. Asian nations like Indonesia top the list of social media users, despite the fact that social media is widely used in America and Europe. As of October 2021, more than 4.5 billion people use social media.

According to Pew Research Center, social media users tend to be younger. Nearly 90% of people between the ages of 18 and 29 used at least one form of social media. People use social media to stay in touch with their friends and extended families. Some people use social media to network for jobs, connect with others around the world who share their interests, and express their ideas, sentiments, insights, and emotions. Participants in these activities are a part of an online social network.

Social networking is a vital resource for businesses. Businesses utilize the platform to locate and interact with customers, increase sales through advertising and promotion, determine consumer trends, and provide assistance or customer care.

The contribution of social media to businesses is substantial. It enables the integration of social interactions on e-commerce sites by facilitating communication with customers. Its capacity for information gathering aids in concentrating marketing and market research activities. In order to attract potential customers, it makes it possible to offer them targeted, timely, and exclusive deals and coupons. Additionally, by linking loyalty programs to social media, social media can aid in developing client relationships.

Benefits: -

The way we all communicate online has changed as a result of social media. It enables us to interact with one another, keep in touch with distant acquaintances, learn about what is happening in the world in real-time, and have access to an enormous quantity of knowledge at our disposal. Social media has in many ways enabled people to connect with one another online and make the world seem more approachable.

At Pew Research Center study found that using social media is associated with having more friends and more diversified personal networks, particularly in emerging nations. For many teenagers, friendships can start virtually, with 57% of teens meeting a friend online.

Businesses are also utilizing social media marketing to target customers directly on their smartphones and laptops, develop a fan base of devoted followers, and establish a culture around their own brands. Some businesses, like Denny's, have developed full Twitter identities in order to sell to younger consumers using their own personas and language.

Drawbacks: -

Individual users may have difficulties with social media in the following ways:

Problems with one's mental health, social media addiction, burnout, and other problems can be brought on by using social media apps excessively.

Polarization. Filter bubbles can contain specific people. When the user is genuinely isolated in an online group that was created using an algorithm, they give the impression that there is open discussion.

Disinformation, polarized environments encourage the spread of disinformation, which is incorrect information that is intended to mislead others. Similar and specific social media difficulties are faced by businesses like offensive posts, intranet and business collaboration discussions have a tendency to stray into other topics. Coworkers may disagree with or find that offensive when that occurs. It might be challenging to regulate these talks and screen out objectionable language.

Protection and preservation, with the functionality offered by collaborative technologies, conventional data protection and retention regulations might not be compatible. Companies may need to cope with increased security threats and compliance difficulties as a result and productivity concerns, social interaction, whether in person or online, might reduce an employee's productivity.

And then finally there's cyberbullying, bullying that takes place online on social media platforms. It uses false social media accounts to psychologically disturb or tease victims. It has its consequences like it affects the user mentally and emotionally making them to feel upset, embarrassed, stupid, angry or even afraid. Which will eventually lead them to developing mental health disorders like PTSD (Post Traumatic Stress Disorder), depression, suicidal thoughts and so on.

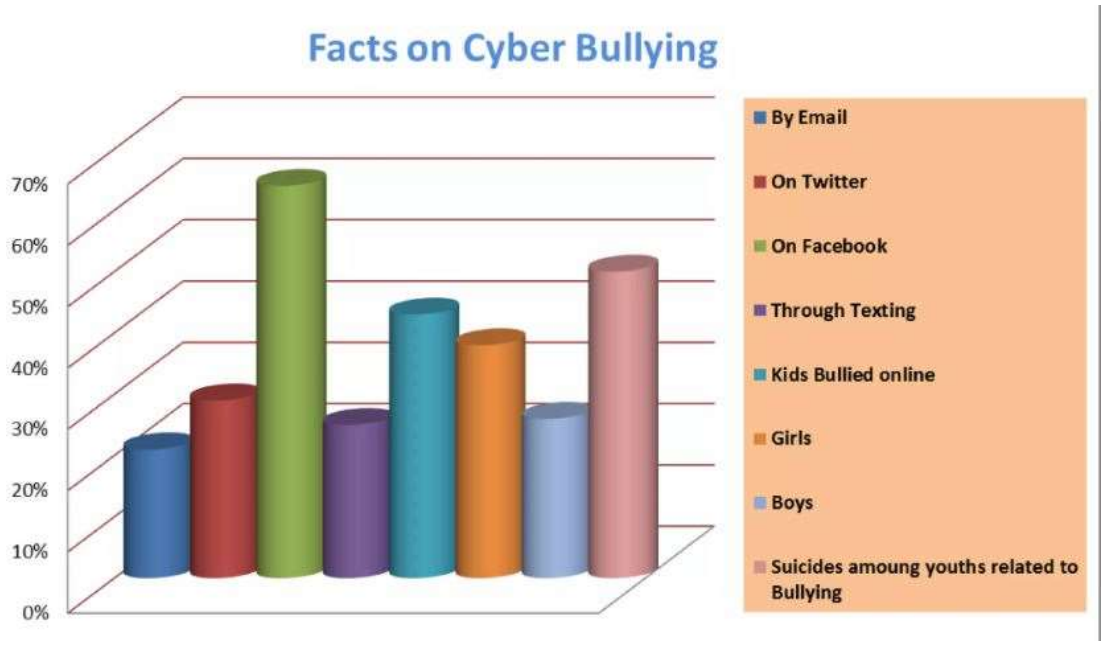

Fig. 1 – Ways of cyberbullying that take place in today's modern world

PTSD is one of the major problems caused due to cyberbullying, so it's a type of anxiety disorder that manifests in response to physical harm or very serious mental or emotional anguish. Its consequences include social disengagement and loneliness, decline in work or academic performance, diminished potential to develop long lasting relationships with others, eating disorders, self-harm, suicidal thoughts and behaviors.

Because of the trauma that is caused to the victim, there is a connection between cyberbullying and depression. After becoming the victim of cyberbullying, a person could experience anxiety and lose interest in previously enjoyed hobbies. In the wake of a cyberbullying incident, depressed episodes may have an impact on one's sleep and eating routines, as well as one's energy levels. In some cases, the victim's memory of the terrible incident may trigger a stress response in the body, leading to PTSD.

2. Literature Survey

In This Section, various methods and techniques are mentioned which include a review of relevant articles published in scientific journals and conference proceedings for detection and repellent of animals in crop fields are explained by considering the different domains.

In [1], The comments on Yahoo! Finance and News were the source of all the data used for training and testing in this study. These comments were monitored by Yahoo staff members, whose primary responsibility was to offer editorial labels for various annotation and editorial activities. In this study, supervised classification using NLP features to assess various user comment characteristics was used. In particular, the Vowpal Wabbit's regression model 5 was used in its default configuration with a bit rate of 28. Prior work in sentiment analysis and text normalization, among other areas, served as the foundation for the usage of NLP features. Four categories can be made from the features: N-grams, linguistic, syntactic, and distributive semantics. To convert some of the data noise that may have an impact on the number of sparse features in the model for the first three features, some light pre-processing is applied. Examples of transformations include normalizing numbers, substituting long unpronounceable phrases with the same token, omitting repetitive punctuation, etc. No normalization was carried out for the fourth feature class.

In [2], The development of a linguistic tool for spotting the early indicators of cyberbullying involved a new approach. Ask.fm, a social network that allows users to post comments and queries to any other user anonymously, served as the place where data was gathered. They begin by automatically labelling each data row. Finally, based on the ratio of received abusive messages, a sliding window is moved through the interaction history of each user to identify probable cyberbullying situations. Finally, since automatic labelling is likely to be noisy, each of these potential cases of cyberbullying is manually annotated to confirm that it contains a cyberbullying incident. To name the non-cyberbullying class, they next use the same procedure. In addition, they train an SVM classifier to detect cyberbullying using a basic collection of lexical, semantic, and aesthetic variables. In this instance, the instances (conversations) are read sequentially in text chunks that are incrementally fed into an SVM classifier to generate the prediction at chunk t. In this instance, we can only construct predictions based on question-answer combinations from the first t chunks of test data at each time t.

In [3], The methods and implications of machine learning (ML) for the automatic identification of cyberbullying have been critically explored in this paper. The selection and discussion of ten literature reviews have been done in this paper. The most popular features in ML models for predicting cyberbullying are those that are content-based. In regard to predicting online bullying from users' online contents, the bag-of-words method is the most popular. Support Vector Machine (SVM), Naive Bayes, and Convolutional Neural Networks were demonstrated to be the most effective algorithms, with SVM being the most effective of the three. Analysis has also been done on the practical ramifications of using these techniques to combat cyberbullying. Online cyberbullying situations can now be quickly and accurately identified because to ML, a revolutionary prevention method. As a result, it could enhance and incorporate initiatives for adolescents.

In [4], They introduced a new framework with the specific goal of capturing sub word information and minimizing the number of architectural factors. Recurrent layers are predicted to keep ordering information even with a single layer; hence their architecture simultaneously incorporates CNN and recurrent neural networks (RNN) on top of unsupervised, pre-trained word vectors. They therefore utilized a recurrent layer in place of the pooling layer in order to more effectively capture long-term dependencies and theoretically prevent the loss of details in local information. On two benchmark datasets, their strategy performed well and outperformed numerous other approaches while achieving a competitive classification accuracy. Their findings showed that a much smaller architecture may do classification at a level comparable to that of larger ones.

In [5], There were several sources used to gather the cyberbullying-related datasets. The information was gathered from a variety of social media sites, including YouTube, Twitter, Wikipedia Talk sections, and Kaggle. Text is included in the statistics, and bullying is indicated or not. Various forms of cyberbullying, including bigotry, hate speech, aggression, insults, and toxicity, are present in the data. After creating the required programs and setting up the cloud environment, it moved on to adopting the common data preparation and cleansing methods for machine learning. This model combines Multinomial Naive Bayes and Optimized Linear SVM, and it achieved accuracy of 92% in both cases. However, accuracy can occasionally lead to overfitting. They consequently chose to perform cross validation and validation. They got the result of an accuracy projecting 96% and a very low mean square error hence their model is certain to have low bias and high variance.

In [6], Jigsaw is used in this instance to gather the data. Also utilized is a dataset of modifications to talk pages on Wikipedia with comments. Comments on Wikipedia are analyzed by Jigsaw. Jigsaw's task is to create and present an analysis technique for personal attacks that blends crowdsourcing and machine learning. The term frequency-inverse document frequency (TF-IDF) technique is used to process text, and this part also includes a quick introduction to text mining. Confusion measures are employed for the model's assessment. The trained dataset, which will distinguish toxic remarks from non-toxic ones, is created using the logistic regression technique. The multi-headed model consists of toxicity (severe toxic, offensive, threatening, insulting, and identity-hating) or Non-Toxicity Assessment, employing confusion measures for their prediction.

In [7], The transfer learning architecture is the model's foundation. In order to address the shared task problem, the USE2 pre-trained model is employed in this study to extract the sentence embedding based on transfer learning architecture. The distributed gradient boosting library (XGB) classifier, XGboost, was created with great efficiency in mind. The USE embedding-powered XGB has been utilized as a text transfer learning model, and it is regarded as a powerful classifier when compared to other machine learning classifiers and deep learning. The aggressive identification for the English language was resolved using this created approach. Before they were ready for the training step using XGB, each input sentence's 512-dimensional pre-trained model had its USE embedding extracted. We adjusted the BERT by removing the final three layers and inserting the Gaussian Noise layer before the GRU in the BERT-GRU training phase. With regard to the BERT-GRU training procedure, we improved the BERT by removing the final three layers and adding a layer of Gaussian noise before a layer of 300 neurons made up of the GRU. Global average pooling, which aims to extract the discriminative features from the previous layer and pass them to the subsequent layer, was then added. Overfitting has been avoided using Dropout and L2 regularization. With a dense layer of one neuron, sigmoid activation function, and truncated normal kernel initializers, the final layer was utilized to anticipate the output predictions.

In [8], This paper focuses on statistics made available by the Wikipedia Detox Project2 (Wulczyn et al., 2017). We concentrate on the Personal Attacks collection out of the three that were released: Aggression and Toxicity. The sentiment lexicon methods, the semantic lexicon approaches, and our baseline model are all described here. As a starting point, we use a convolutional neural network (CNN). Furthermore, CNN outperformed other algorithms in early tests on our data (e.g., RNN or GRU). The model's input is a word embedding matrix that is first generated after words are extracted from a corpus. As demonstrated by their experiments, using semantic lexicons to improve word embeddings holds promise for the job of abusive language detection. These techniques are quick and flexible and, in addition to increasing detection accuracy for our data (in the form of F1-macro), they do not change or rely on the type of learning model.

In [9], a method of detecting cyberbullying that is objective and bases its judgements on the semantics of a social media conversation rather than delicate cues that could be indicators of cyberbullying, like the words "gay," "black," or "fat." The False Negative Equality Difference (FNED) and False Positive Equality Difference (FPED) metrics can measure bias in a text classification algorithm. In this study, they looked at unintentional biases in databases for session-based cyberbullying detection. Social media sessions are made up of a string of remarks with rich contextual formation, in opposition to conventional data for bias mitigation in text classification. A successful debiasing approach has been suggested by utilizing RL methods in order to reduce these unintended biases.

In [10], A necessary step in understanding how a person's role changes over time is to identify participants' bullying roles. We add two new roles—reporter (who might not be present during the occurrence, unlike a bystander) and accuser—to the traditional role structure to better address bullying footprints in social media (accusing someone as the bully). Due to a lack of detail in the tweet, both positions could be victims, defenders, or bystanders in the traditional sense. There are four roles that appear most frequently in social media: accuser (A), bully (B), reporter (R), and victim (V). In the subsequent investigation, we combined every last role into a broad category called "other" (O). So they have listed some important issues as NLP tasks after identifying them.

In [11], This research looked at how often it is for hate speech to spread on social media and the different strategies put out in the literature to stop it. A hybrid deep learning model that can detect and block hate speech on social media platforms has been effectively constructed in this research based on the gaps in the literature that were found (Twitter and Facebook). In this study, they tested a number of deep learning models (RNN, CNN, and hybrid CNN-LSTM) for detecting hate speech online. They came to the conclusion that hate speech may be identified and suppressed on social media sites before it reaches the general audience. In order to reduce the threat of online hate speech, this report advises Twitter and Facebook to take the findings into consideration.

In [12], They suggested a traditional technique for identifying cyberbullying in the Instagram text comments dataset. Their planned work's primary contribution is as follows: To employ Fast Text and Similarity measures to construct a lexical-based conventional model for analysing the morphology and word order of the harassing words in text comments. to identify the intended audience and motivation behind a specific textual comment that was recovered by NLP utilising the feature extraction method. Using a quick text supervised approach in conjunction with a word similarity metric, the loss function and time complexity of the detection model are reduced. Intention model outperforms J48, NB, SVM, RF, bi-LSTM, and MLP neural networks, it is concluded. By adding fast text, the time complexity of the model is lesser due to memory management.

In [13], Social networking and other internet-based applications are becoming associated with fear and negativity thanks to cyberstalkers and other online criminals. In this study, six machine learning classifiers—Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT), k-Nearest Neighbor (KNN), and Naive Bayes—were used to implement several popular features extraction methods. The goal was to detect cyberstalking (NB). Based on experimental findings, it was determined that simple feature extraction approaches performed better than sophisticated feature extraction methods. The findings of these classifiers are nearly identical, and SVM, RF, and LR classifiers outperformed all deployed feature extraction techniques. According to experimental findings, the detection model's performance was impacted by the feature extraction techniques.

In [14], This study determines mental diseases using Reddit.com user data that was suggested by Murarka and Radhakrishnan. To automatically identify mental illnesses in social media texts, they used conventional machine learning, deep learning, and transfer learning techniques. In this document, six mental disorders are listed: none, PTSD (Post Traumatic Stress Disorder), bipolar disorder, depression, anxiety, and ADHD (Attention Deficit Hyperactivity Disorder). They started off by using four distinct machine learning classifiers: Random Forest, Linear Support Vector Machine, Multinomial Naive Bayes, and Logistic Regression. Modern neural network models were used in the second type of technique, which is deep learning-based. They used a number of pre-trained deep learning algorithms for multi-class mental illness identification, including Convolutional Neural Network, Gated recurrent unit (GRU), Bidirectional Gated recurrent units (Bi-GRU), and LSTM.

In [15], With the purpose of automating the identification and management of cyberbullies on social media platforms, the authors of this research have suggested a model that serves as a prototype for a cyberbullying detection system. This concept utilizes real-time data, and the generated web portal would resemble a copy of social media websites like Twitter, where users may publish tweets and have those changes reflected in the feed. By putting the tweeted tweet through their model, this website will determine whether it contains any cyberbullying content. Before to feeding training data into stacked word embeddings, the data is cleaned and preprocessed. The CNN-BiLSTM deep learning model is subsequently trained to outperform other trained deep learning models. It is concluded that CNN-BiLSTM model has the best accuracy. While the CNN alone can only train local characteristics from word n-grams, with its LSTM layer, the CNN-BiLSTM can also learn global features and long-term dependencies.

In [16], They have put up a strategy to identify cyberbullying using machine learning methods. They employed TFIDF and sentiment analysis techniques to extract features, and they tested their model on two classifiers: SVM and neural networks. Several n-gram language models were used to evaluate the classifications. Using the Kaggle cyberbullying dataset, they assessed the suggested approach. When combining TFIDF and sentiment analysis, they

achieved 90.3% accuracy with SVM with 4-grams and 92.8% accuracy with Neural Network with 3-grams. Since their neural network likewise obtains an average f-score of 91.9%, compared to the SVM classifier's f-score of 89.8%, they came to the conclusion that their neural network performed better.

In [17], The experimental research in this paper demonstrated that DEA-RNN had outperformed all other available approaches in all the scenarios with regard to a variety of metrics, including precision, specificity, recall, accuracy, and F-measure. This represents the effect of DEA on RNN performance. The performance rates of their hybrid suggested model were higher than those of the other models that were taken into consideration, but when the input data is increased beyond the initial input, the DEA-feature RNN's compatibility decreases. As a result, their suggested DEA-RNN is likewise guaranteed to be very flexible for current specialised short text topic models.

In [18], In order to categorize tweets as cyberbullying or non-cyberbullying and their severity in Twitter as low, medium, high, or none, researchers created a feature-based model that incorporates features from tweets' contents. They used PMI-semantic orientation in addition to the Embedding, Sentiment, and Lexicon characteristics. Naive Bayes, KNN, Decision Tree, Random Forest, and Support Vector Machine methods were used to process the extracted features. They employed a publicly accessible lexicon and annotated dataset from a git repository for their investigation. As a result of how well the Basic classifier managed the distribution of class imbalance, it can be stated that its overall performance marginally increased when the SMOTE setting was activated. SMOTE and all other characteristics, as well as the results for true positives and false positives rate for each classifier, showed a considerable improvement in performance in terms of Kappa, F-measure, and accuracy. Displays that, when compared to other classifiers, Random Forest has the highest true positive rate of 91% and the lowest false positive rate (29%).

In [19], Their paper analyzed the body of research on the topic of detecting aggressive conduct on social media platforms using machine learning techniques, specifically Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbor (KNN). They especially looked at four elements—data collecting, feature engineering, building of a cyberbullying detection model, and evaluation of built-in cyberbullying detection models—for identifying messages of cyberbullying using machine learning techniques. Researchers came to the conclusion that cyberbullying was accurately detected by the SVM-based algorithm. According to the findings, the rule-based Jrip cyberbullying model is more precise than the SVM-based model, although it is less dependable. Yet, compared to NB and tree-based J48, the SVM-based cyberbullying model is more precise. Data from Twitter was used to build an SVM-based model of cyberbullying.

In [20], They explained how to use a novel FSSDL-CBDC method for identifying and categorizing online bullying. They proposed the FSSDL-CBDC technique, which comprises several phases, including pre-processing, feature selection, and classification, among others. The BCO-FSS approach was developed to select the ideal group of features from the pre-processed data, and it also considerably improves the classification outcomes overall. Similar to the previous models, the SSA-DBN model also receives and categorizes the feature-reduced subset. By employing the SSA to adjust the DBN model's hyperparameter in comparison to the traditional DBN model, better results were obtained. The increased detection performance of the suggested FSSDL-CBDC technique, which was successful, was evaluated using a large number of simulations on a benchmark dataset. The simulation results showed that the FSSDL-CBDC strategy performed much better in classification than the others when compared to other cutting-edge methods.

3. Implications of Our Findings

These strategies and methodologies have the potential to increase the reliability and promptness of cyberbully identification, which can lessen the harm done to people, especially teens, and usher in a new era of proper natural language processing in real-time processing. Using AI technology can make it possible to take detection actions that are more precise and effective, hence reducing the negative effects on individuals and their environments. We have researched deep learning, machine learning, and other learning-based methodologies from these research and conference publications. For some businesses, adopting these strategies and tactics could necessitate sizable infrastructure and technological investments. Due to the gathering and analysis of user data, such as posts and other communications, using these technologies may also give rise to privacy and security concerns.

More study is required to identify the most successful strategies in various circumstances because the efficacy of these tactics and procedures may change based on the particular circumstance. It may also be necessary for users, developers, and other stakeholders to undergo new training, which could be difficult in some circumstances. The models performed well when compared to the experimental results, but their accuracy was limited by a number of factors, such as the lack of a large dataset, real-time processing, the inability to comprehend slang, the context of sentences, and the models' limited capacity to recognize ever-evolving slangs, vocabulary, and other issues. The use of AI in cyberbully detection will be revolutionary in the ensuing years and will have wider ramifications for the NLP sector, including potential effects on employment and the adoption of other technologies.

With these things said we found out some differences from each paper and their approaches to either use ML or DL so we have mentioned a few differences below.

Machine Learning	Deep Learning
Only a little amount of training data is needed.	A significant amount of training data is needed.
The characteristics that the system will use must be manually identified by the coder or user.	Based on the training data, the machine automatically recognises the important features.
Most of the time, attention is given to resolving a specific issue.	Focus is typically placed on locating intriguing patterns in the training dataset.
Each component of the problem is broken down, solved separately, and then integrated.	An end-to-end processing technique is used throughout.

Given the explicit guidelines that are supplied, the output result is simple to grasp.	Due to the system's use of its own logic to make decisions, the reasoning may be challenging to comprehend.
Tests take longer to complete.	Test take less time to complete.
Even with less computing power, it can function.	It requires a lot of processing power.
Using historical data, machines are capable of autonomous decision-making.	Artificial neural networks aid in the decision-making of machines.

4. Conclusion

We have examined a number of strategies and methods for identifying and stopping cyberbullying. Deep learning, machine learning algorithms, and other technologies are used in these strategies and procedures to find texts that are regarded to be violent. Through the provision of more precise and timely information on the material that the poster posts as well as the facilitation of more focused and effective preventative measures, the use of AI technology has the potential to increase the effectiveness of cyberbully identification. The use of these technologies may, however, come with difficulties, such as the requirement for substantial investments and the acquisition of new knowledge and expertise. The best methods for utilizing AI technologies in cyberbully detection require more study. Further research is needed to determine the optimal approaches for using AI technologies in cyberbully detection, and to assess the potential impacts on the social media platforms and other stakeholders.

Therefore, based on what we have discovered through our research, cyberbullying detection still needs a lot of effort and training, and given the limitations of machine learning, we need to explore more options in deep learning. DL offers a variety of ways and methods to handle and train large amounts of data, build context aware models, detect emoji along with text, and other capabilities that would be made possible with further research and a deep dive into DL.

References

[1] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. (2016). Abusive Language Detection in Online User Content. International World Wide Web Conference Comittee (IW3C2).

[2] Niloofar Safi Samghbadi, A. Paastor Lopez-Monroy, and Thamar Solorio. (2020). Detecting Early Signs of Cyberbullying in Social Media. Proceedings of the Second Workshop on Trolling, Aggression, and Cyberbullying (LREC 2020).

[3] Jacopo De Angelis, and Giulia Perasso. (2020). Cyberbullying Detection Through Machine Learning: Can Technology Help to Prevent Internet Bullying?. International Journal of Management and Humanities (IJMH), Volume-4 Issue-11, July, 2020.

[4] Abdalraouf Hassan, (Member, IEEE), and Ausif Mahmood, (Senior Member, IEEE). (2018). Convolution Recurrent Deep Learning Model for Sentence Classfication. IEEE Access. Digital Object Identifier 10.1109/ACCESS.2018.2814818.

[5] Tosin Ige, and Sikiru Adewale. (2022). AI Powered Anti-Cyber Bullying System using Machine Learning Algorithm of Multinomial Naive Bayes and Optimized Linear Support Vector Machine. International Journal of Advanced Computer Science and Applications, Vol. 12, No. 5, 2022.

[6] P. A. Ozoh, A. A. Adigun, and M. O. Olayiwola. (2019). Identification and Classification of Toxic Comments on Social Media using Machine Learning Techniques. International Journal of Research and Innovation in Applied Science (IJRIAS).

[7] Saja Khaled Tawalbeh, Mahmoud Hammad, and Mohammad AL-Smadi. (2020). SAJA at TRAC 2020 Shared Task: Transfer Learning for Aggressive Identification with XGBoost. Proceedings of the Second Workshop on Trolling, Aggression, and Cyberbullying (LREC 2020).

[8] Anna Koufakou, and Jason Scott. (2020). Lexicon-Enhancement of Embedding-based Approaches Towards the Detection of Abusive Language. Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (LREC 2020).

[9] Lu Cheng, Ahmadreza Mosallanezhad, Yasin N. Silva, Deborah L. Hill, and Huan Liu. Mitigating Bias in Session-based Cyberbulyying Detection: A Non-Compromising Approach. (This work is based upon work supported by the NSF Grants 1719722 and 2036127).

[10] Jun-Ming Xu,Kwang-Sung Jun,Xiaojin Zhu, and Amy Bellmore. (2012). Learning From Bullying Traces in Social Media. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[11] Hyellamada Simon, Benson Yusuf Baha, and Etemi Joshua Garba. (2022). A Multi-Platform Approach Using Hybrid Deep Learning Models For Automatic Detection Of Hate Speech On Social Media. Bima Journal of Science and Technology, Vol. 6 (2) Aug, 2022.

[12] S. Abarna, J.I. Sheeba, S. Jayasrilakshmi, S. Pradeep Devaneyan. (2022). Identification of cyber harassment and intention of target userson social media platforms. Engineering Applications of Artificial Intelligence 115 (2022) 105283.

[13] Arvind Kumar Gautam and Abhishek Bansal. (2022). Effects of Features Extraction Techniques on Cyberstalking Detection Using Macine Learning Framework. Journal Of Advances in Information Technology Vol. 13, No. 5, October 2022.

[14] Iqra Ameer, Muhammad Arif, Helena Gomez-Adorno, Grigori Sidorov, and Alexander Gelbukh. (2022). Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning. arXiv:2207.01012v1 [cs.LG]

[15] Mitushi Raj, Samridhi Singh, Kanishka Solanki, and Ramani Selvanambi. (2022). An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques. SN Compter Science, A Springer Nature Journal.

[16] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyed Emad, Eslam Amer, and Ammar Mohammed. (2019). Social Media Cyberbullying Detection using Machine Learning. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 5, 2019.

[17] Belal Abdullah Hezam Murshed, Jemal Abawajy, (senior Member,IEEE), Suresha Mallappa, Mufeed Ahmad Naji Saif, and Hasib Daowd Esmail Al-Ariki. (2022). DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform. B. A. H. Murshed et al. IEEE Access. Digital Object Identifier 10.1109/ACCESS.2020.3153675.

[18] Bandeh Ali Talpur, Declan O'Sullivan. Cyberbullying Severity detection: A machine learning approach. (2020). PLOS ONE 15(10).

[19] Mohammed Ali Al-Garadi, Mohammed Rashid Hussain, Nawsher Khan, Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, and Abdullah Gani. V. (2019). Predicting Cuberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. M. A. Al-Garadi et al. IEEE Access. Digital Object Identifier 10.1109/ACCESS.2019.2918354.

[20] Neelakandan S, Sridevi M, Saravanan Chandrasekaran, Murugeswari K, Aditya Kumar Singh Pundir, Sridevi R, and T Bheema Lingaiah. (2022). Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media. Computational Intelligence and Neuroscince Volume 2022, Article ID 2163458.