International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 10 | Oct 2023 www.irjet.net p-ISSN: 2395-0072

BINARY TEXT CLASSIFICATION OF CYBER HARASSMENT USING DEEP LEARNING

Srinadh Unnava[1], Manikanta Bandla[2]

[1] *Assistant Professor dept. of Information Technology SITE,Tadepalligudem srinadh@sasi.ac.in*

[2] *Student dept. of Information Technology SITE,Tadepalligudem manikanta1208@sasi.ac.in*

Abstract -

Social media platforms such as Facebook and Twitter offer various benefits, but they also have some negative aspects. One of the significant issues related to these platforms is cyberbullying, which can have a profound impact on the victims. The impact of cyberbullying may vary from person to person, and it is subjective to how each individual deals with it.The influence of cyberbullying on youths has become more prominent in recent times. With the aid of machine learning, it is possible to identify speech patterns employed by bullies and their targets, and establish guidelines to automatically recognize cyberbullying material. It affects both young people and adults and has been linked to negative outcomes such as depression and even suicide.To successfully address cyberbullying, it is necessary to first comprehend the numerous elements and processes that lead to its occurrence. Furthermore, there is a rising awareness of the necessity to govern material uploaded on social media networks. The primary goal of this research is to create a cyberharassment detection system (CDS) to detect harassing and abusive activity on electronic media platforms.Among such algorithms, we investigated four distinct optimizers of neural networks and convolutional neural networks in this study. Rmsprop had the highest accuracy (98.45%), followed by Adam, Adadelta, and Adagrad. Index Terms—Deeplearning, Cyberharass, Cyberbullying, Machinelearning, Electrionic media,twitter.

Key Words: *Deeplearning, Cyberharass, Cyberbullying, Machine Learning, Electrionic media,twitter.*

1.INTRODUCTION

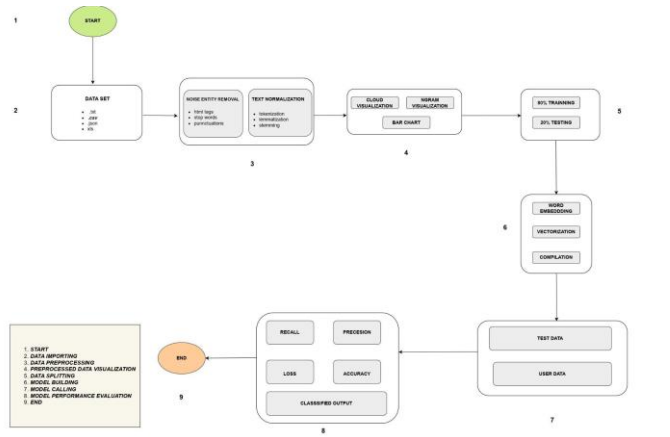
The growth of information and communication technology has provided numerous advantages to society, such as social media, which allows people to expand their social networks. However, as Segal suggests, this new technology has a dark side, including cyberbullying. Cyberbullying is a form of online violence where the victim is humiliated, ridiculed, and intimidated. This kind of harassment can lead to serious mental health issues, and in some cases, even suicide. Social media platforms offer users the opportunity to express their opinions and connect with people around the world. However, the two sided nature of these platforms has contributed to anti-social behavior that affects people of all ages. Cyberbullying and cyber-aggression are prevalent, with over half of young social media users experiencing digital harassment.

2. RELATED WORKS

Roy, et al [1] proposed a method that can prevent image based cyberbullying on social media platforms. To accomplish this, the researchers first used a deep learning-based convolutional neural network to build the model. However, they later turned to transfer learning models, which involve reusing pre-trained models, in order to improve the model's performance. After testing various settings of the hyper-parameters, the researchers found that the transfer learning-based model was more effective in addressing this problem. This research demonstrates the importance of developing new tools and technologies to combat the harmful effects of cyberbullying on individuals and communities. Aldhyani, et al [2] created a Cyberbullying Detection System (CDS) that can identify harmful and abusive behavior on social media platforms. The researchers conducted two experiments to train and test the proposed system using binary and multiclass cyberbullying classification datasets. This research highlights the importance of developing advanced tools and technologies to address the growing problem of cyberbullying on social media platforms. Ahmed, et al [3] introduces a binary and multiclass classification model that utilizes a hybrid neural network to detect instances of bullying in the Bengali language. The researchers used a dataset of 44,001 user comments from popular public Facebook pages, This research highlights the importance of developing effective models for identifying and preventing instances of bullying in online environments, especially in languages other than English. De Angelis, et al [4] said that with the rise of technology, cyberbullying has become a significant threat to the psychosocial wellbeing of adolescents. ML encompasses a wide range of techniques that allow systems to learn from data and make complex decisions quickly. This paper aims to explore the potential of ML in detecting and preventing cyberbullying, highlighting the importance of developing effective tools and strategies to address this growing problem. Keni, et al [5] This study focuses on developing techniques for detecting cyberbullying using supervised learning methods. This research highlights the potential of

machine learning techniques to aid in the detection and prevention of cyberbullying, which is an important step towards creating safer online spaces for all users. Chandrasekaran,et al [6] says Because of the ease of access to the internet and cell phones, the popularity of online social networks and social media has skyrocketed. However, security and privacy issues are key concerns on these platforms, and cyberbullying has developed into a serious issue that must be addressed. Cyberbullying is defined as the repeated and intentional use of information and communication technology platforms such as social media and the internet to transmit hate messages.Furthermore, the SSA-DBN model combines the salp swarm algorithm (SSA) with a deep belief network (DBN) to detect and characterise cyberbullying in social media networks and other online contexts. The creation of the BCO-FSS and SSA-DBN models for the identification and categorization of cyberbullying is a first in the area. Numerous simulations show that the suggested FSSDL-CBDC approach has higher classification performance. Aditya, et al [7] recommended the construction of a model for detecting cyberbullying that takes several factors into account. We used BERT, a bidirectional deep learning model, to achieve some of these properties. Talpur, et al [8] article presents a thorough analysis of cyberbullying that commonly takes place on online social networking (OSN) platforms. It offers an overview of the existing techniques to address cyberbullying on OSNs and examines the obstacles and difficulties involved in the development of cyberbullying detection systems. The paper also suggests future research directions in this area. The paper begins by introducing OSNs and cyberbullying, outlining different forms of cyberbullying, and discussing the accessibility of data. This project aims to examine and assess the interaction between two individuals or an anonymous user with the help of a machine learning model. The primary objective is to detect and prevent harassment in one-on-one conversations. Two classifiers, SVM and Na¨ıve Bayes, are employed to train and test the detection of cyberbullying in social media [9].The nature of online social networks provides an opportunity for cyberbullies to target people across different regions and countries. Our goal is to detect cyberbullying in Twitter using SVM. Our objectives are outlined in the objective section. In this paper they categorizes existing methods into four main classes: supervised learning, lexicon-based, rule-based, and mixed-initiative approaches. We will focus on supervised learning-based approaches that utilize classifiers like SVM and Na¨ıve Bayes to create predictive models for cyberbullying detection. We will also employ natural language processing techniques and machine learning methods such as Bayesian logistic regression, random forest algorithm, and support vector machines to identify cyberbullying [10]. In today's fast-paced world, innovation is rapidly increasing, leading to increased communication. Unfortunately, this communication can also be used for negative purposes, such as cyberbullying. Cyberbullying involves targeting individuals or groups through electronic messages on social media platforms, which can lead to depression and even suicide. In fact, approximately 80 Percent of young people who commit suicide have experienced depression caused by cyberbullying. In this paper they discusses the various approaches and ideas that can be used to detect cyberbullying on social media platforms, as well as its impact on individuals [11].The emergence of social networking sites on the internet has facilitated diverse expressions and opinions. This study proposes a sentiment detection system using text and image data. The text data is obtained from Twitter using its API, and RNN and CNN are used for text and image analysis respectively. The models' efficiency and performance were evaluated, and RNN and CNN were found to outperform other classification algorithms [12]. Research has indicated that cyberbullying is a prevalent issue across the globe, affecting both young individuals and adults. Regulating social media content has become increasingly necessary. The upcoming study aims to analyze hate speech tweets from Twitter and personal attack-based comments from Wikipedia forums to extract text data utilizing natural language processing and machine learning. The objective is to create a model that can detect cyberbullying. The study considers three feature extraction methods and four classifiers to identify the optimal approach. [13].Gauri, et al [14] analysed the current cyberbullying detection models and proposes a novel approach utilizing Na¨ıve Bayes Classification and N-Gram Model to evaluate the overall bullying scenario or sentiment of tweets. Monirah Abdullah Al-Ajlan and Mourad Ykhlef [15] aimed to improve cyberbullying detection techniques by proposing a new algorithm called CNN-CB. Unlike traditional approaches, CNN-CB does not require feature engineering and produces better predictions. The algorithm is based on word embedding, where similar words have similar representations, and is implemented using a convolutional neural network (CNN) that incorporates semantics. In our experiments, CNN-CB outperformed traditional content-based cyberbullying detection methods. The task of sentiment analysis is a complex one that falls under the domain of Natural Language Processing (NLP), text analytics, and computational linguistics. In this research paper,they suggested a supervised machine learning method for detecting and preventing cyberbullying [16]. The performance of eleven classification models, including four standard machine learning algorithms and seven shallow neural networks, is compared using a unique neural network design in which parameters are optimised. The researchers used two real-world cyberbullying datasets to evaluate the effects of feature extraction and natural language processing based on word embeddings. According to the results, bidirectional neural networks and attention models provide great accuracy, whereas logistic regression is the best traditional machine learning classifier. Among the neural networks tested,The research examines eleven classification approaches and seven feature extraction methodologies, and the shallow neural networks proposed outperform existing state-of-the-art methods for identifying cyberbullying [17].

3. PROPOSED DETECTION MODEL



1.1 Dataset Description

In this experiment dataset was considered from which randomly selected 9093 tweets to focus on identifying cyberbullying. These data are saved as a CSV file with two columns: 'tweet' and 'class'."Tweet denotes the collection of all offensive and non-offensive tweets obtained for our experiment. Class represents how each tweet is labelled. Tweets that are not offensive are labelled as '0,' while those that are offensive are labelled as '1'. The remaining 4929 tweets are classified as offensive (0.542), while 4164 are classified as non-offensive (0.458). Many preparation procedures were used to obtain the data in a clean way. The technique included noise reduction (tags, links, whitespace, and numbers), punctuation removal, and conversion of the entire document to lower case for uniformity. The documents are then tokenized, and insignificant terms known as stopwords are removed. the tokens are

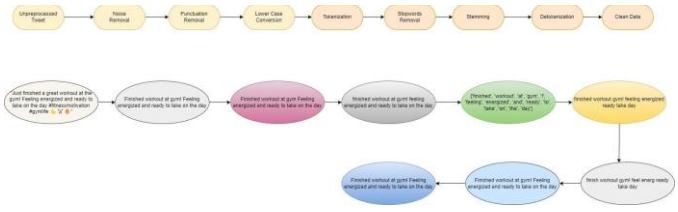


Fig.2 : Data preprocessing flow.

transformed to their root form, a process known as stemming. Snowball Stemmer was employed for this purpose.Finally, in the final phase of preprocessing, detokenization was performed, resulting in clean and acceptable data for the following stage.

1.2 Preprocessed Data Visualization

Data that has been preprocessed The process of visualising data after it has been cleaned, converted, and readied for analysis is referred to as visualisation." This form of visualisation is significant because it helps us examine and comprehend the data's underlying patterns and insights without being distracted by noise or extraneous details.

1.3 Data splitting

The cleaned label dataset was divided into two parts: a train set (80%) and a test set (20%). Following the completion of the division, features were added to the train set. Because our dataset was not completely balanced, we verified our model using a variety of cross-validation approaches, including KFold, Stratified K-Fold, Shuffle Split, and Stratified Shuffle Split. These procedures have been applied tenfold.

1.4 Model building

1.4.1 NEURAL NETWORK

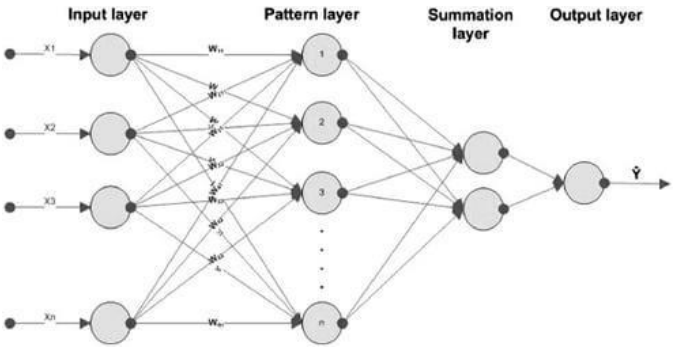


Fig.3 : Artificial Nueral Network..

ANNs, also known as neural networks (NNs), are computer systems inspired by biological neural networks in animal brains. They are made up of linked components known as artificial neurons .Signals can be sent between neurons via connections, also known as edges. Each artificial neuron receives signals, analyses them, and delivers them to other neurons that are linked to it. The weights of the connections between neurons alter when learning happens, increasing or lowering the intensity of messagesy

1.4.1.1 Adam optimizer

Adam optimizer: Adaptive Moment Estimation (Adam) is a powerful optimisation approach that is utilised in gradient descent. It is especially beneficial for dealing with large-scale challenges involving enormous volumes of data or parameters.

1.4.1 .2 Adagrad optimizer

Adagrad is an abbreviation for Adaptive Gradient Optimizer. Optimizers such as gradient descent, stochastic gradient descent, and mini-batch SGD were utilised to lower the loss function in relation to the weights.In SGD and mini-batch SGD, the value of used to be the constant for each weightage, or say for every parameter. Here typically = 0.01. Where in Adagrad Optimizer, the basic idea was that each weight has a different learning rate (). This change is critical because in a real dataset, some features are sparse (for example, in Bag of Words, most

features are zero, so it is sparse) and some are dense (most features will be non-zero)

1.4.2 CONVOLUTIONAL NEURAL NETWORK,

A convolutional neural network (CNN) is a sort of artificial neural network that is often used for visual data analysis. Unlike classic fully connected networks, CNNs include a mathematical technique known as convolution in at least one of their layers, allowing them to exploit data's hierarchical structure. CNNs divide the input into smaller, CNNs are modelled after

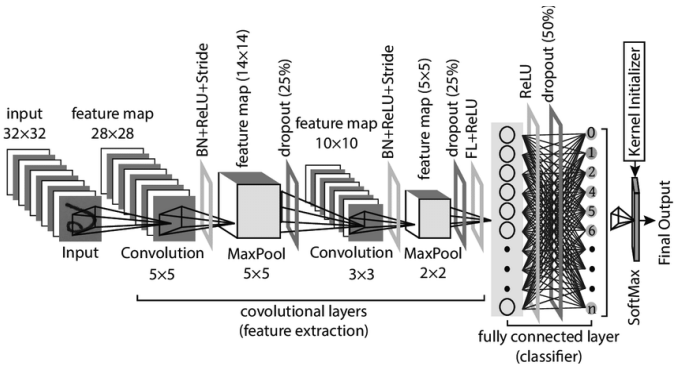


Fig.4 : Convolutional Nueral Network

Convolutional Nueral Network the organisation of the visual cortex in animals, in which individual neurons react exclusively to stimuli in a narrow part of the visual field known as the receptive field.

ARCHITECTURE:

The very next layer from the paste layer (if any should occur) is a convolutional layer. This is the main core layer of the entire convolutional neural network.In this paper different convolution layers are used. follows:

- Input.
- Word Embedding layer.
- Dropout layer(0.25).
- Convolution layer.
- Max Pooling.
- Convolution-layer.
- Max Pooling.
- Dropout layer(0.5).
- Softmax.
- Classification layer

1.4.2 .1 Rmsprop optimizer

RMSprop is a technique used in optimizing neural networks during training. It was introduced by Geoffrey Hinton, the creator of back-propagation. In complex functions like neural networks, gradients tend to vanish or explode as data passes through the function. wt+1 = wt – αt (vt+ε) 1/2 ∗ h δL δwt i Essentially, RMSprop adjusts the learning rate dynamically instead of treating it as a fixed hyperparameter. As a result, the learning rate changes over time

1.4.2 .2 Adadelta optimizer

Adadelta is another AdaGrad variation. The primary distinction is that it decreases the extent to which the learning rate is adaptable to the coordinates. It has also been described as having no learning rate since it uses the amount of change as a calibration for future adjustments. wt = wt−1 − η ∂L ∂w(t−1) Zeiler (2012) proposed the algorithm. Given the preceding algorithm's description, this is quite basic.

1.4 Model Training

Model training is a vital stage in the development of a deep learning-based text classification system. The goal is to lower the loss function on a labelled training dataset by optimising the model's parameters. The procedure entails randomly initialising model parameters, feeding input data into the model to compute a predicted output, evaluating the difference between expected and actual output, and repeating the process

4. EXPERIMENTAL RESULTS

The present research covers multiple DL outcomes from experiments and transfer learning models.The dataset statistics are used for testing and training. Accuracy is the total number of correctly categorised samples among all samples, while recall is the number of genuine positive samples over all positive samples. Precision is the number of real positive samples over all samples shown to be positive. The total dataset size is 8000, with 3542 samples classified as non-bully (Class 0) and the remaining 4458 samples classified as bullies (Class 1).

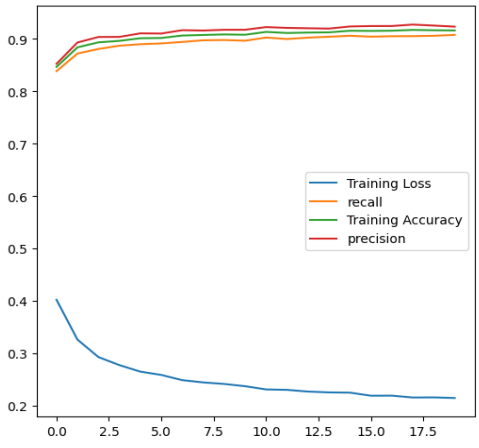


Fig.5 : Adam optimizer Performance.

The model, on the other hand, has experimented with various training and testing sizes of samples. Because the suggested system is a supervised model, its performance is measured using precision, recall, loss, confusion matrix, area under the ROC curve, and accuracy.

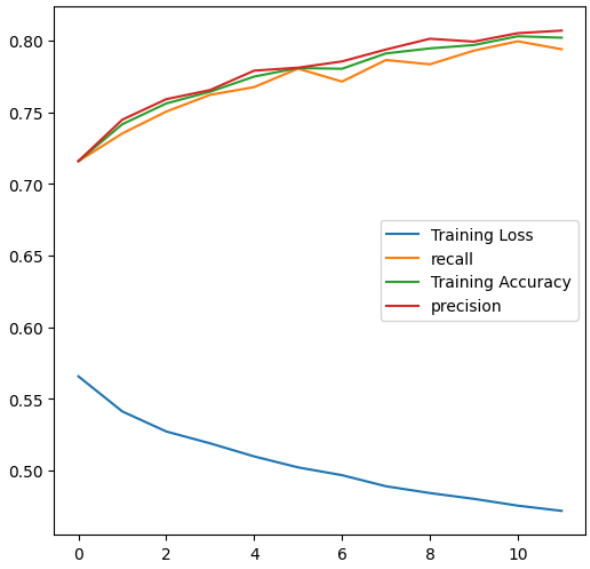


Fig.6 : Adagrad optimizer Performance.

Precision is the number of properly categorised bullying comments among all bullying comments in the dataset, whereas recall is the number of bullying comments among all bullying comments in the dataset. The loss is the mean error across samples for each update (batch) or the average error across all updates for the samples (epoch). The total of properly categorised bullying and non-bullying remarks is the accuracy.The area under the ROC curve (AUC) has a value that ranges from 0 to 1.

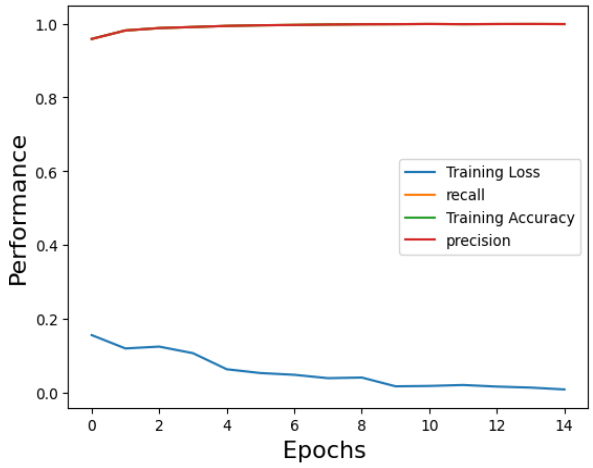


Fig.7 : Rmsprop optimizer Performance.

The Table displays the input processing and prediction results obtained throughout our testing.

Table -1: Model performances

	Accuracy	Loss	Precession	Recall
Adam	0.9173	0.216	0.9274	0.905
Adagrad	0.8032	0.4755	0.8054	0.7997
Adadelta	0.8032	0.2905	0.8739	0.8739
Rmpsprop	0.8032	0.1178	0.9845	0.9845

We took a bullyingrelated tweet from Twitter and applied it to our model. The categorization report of Adam Optimizer is displayed in Fig. 5. Adagrad optimizer's categorization report is displayed in Fig. 6.The Adadelta optimizer's categorization report is displayed in Fig. 7.

The classification report of the Rmsprop optimizer is displayed in Fig. 8. Bullying and non-bullying are represented here by the designations 0 and 1, respectively. Based on the results of this research tests, the performance is depicted in Fig. 5.

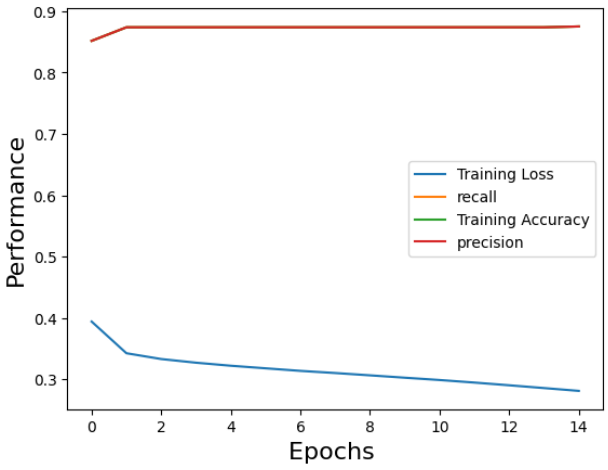


Fig.8 : Adadelta optimizer Performance.

The accuracy of the Adam, Adagrad, Adadelta, and Rmsprop is represented in Table 1 and is 91.73%, 80.32%, 87.39%, and 98.45%, respectively. The loss of the Adam, Adagrad, Adadelta, and Rmsprop is parallel represented in Table 1 and is equal to 21.60%, 47.55%, 11.78%, and 29.05%, respectively. When applied to the Twitter dataset for the cyberbully classification, the Rmsprop optimizer method outperformed the other proposed models with a better accuracy of 98.45%, which may be viewed as a better outcome when compared to the typical machine learning models and optimizers employed on comparable datasets.

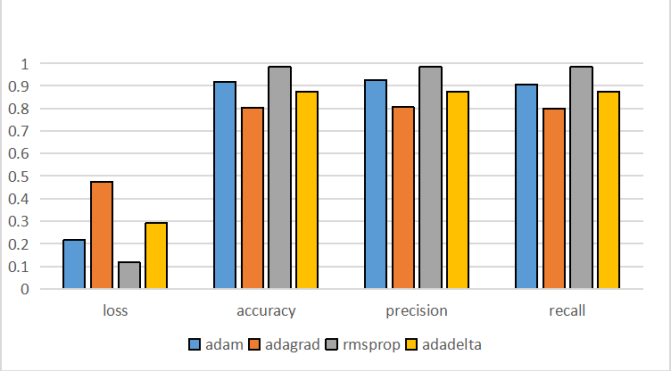


Fig.9 : Optimizers Performance.

Graph shows the overall performances of Adam, Adadelta, Adagram, Rmsprop optimizers.

5. CONCLUSION AND FUTURE WORKS

Online harassment occurs when someone threatens another person electronically using social media, online discussion blogs, email, or other electronic or digital means. The goal of this research is to design and build a cyber harassment detection system that can be utilised to detect and eliminate online bullying occurrences perpetrated by social media users. Create a deep learning approach for detecting abusive tweets and chats online prior to cyberbullying, which may be employed in the creation of bullying detection systems for online social media sharing platforms like Twitter and Facebook. When compared to other optimizers, the Rmsprop optimizer algorithm performed better. In future, significance of individual features can be studied for further enhancement of the model.

REFERENCES

[1] Roy, P.K., Mali, F.U. Cyberbullying detection using deep transfer learning. Complex Intell. Syst. 8, 5449–5467 (2022). https://doi.org/10.1007/s40747-022-00772-z .

[2] Aldhyani, T.H.H.; Al-Adhaileh, M.H.; Alsubari, S.N. Cyberbullying Identification System Based Deep Learning Algorithms. Electronics 2022, 11, 3273. https://doi.org/10.3390/electronics11203273.

[3] Ahmed, Md Faisal, et al. "Cyberbullying detection using deep neural network from social media comments in bangla language." arXiv preprint arXiv:2106.04506 (2021).

[4] De Angelis, Jacopo, and Giulia Perasso. "Cyberbullying detection through machine learning: Can technology help to prevent internet bullying?." International Journal of Management and Humanities 4.57 (2020): 10-35940.

[5] Keni, A., et al. "Cyberbullying detection using machine learning algorithms." Int. J. Creat. Res. Thoughts (IJCRT) 1972 (1966): 2020.

[6] Chandrasekaran, Saravanan, Aditya Kumar Singh Pundir, and T. Bheema Lingaiah. "Deep learning approaches for cyberbullying detection and classification on social media." Computational Intelligence and Neuroscience 2022 (2022).

[7] Desai, Aditya, et al. "Cyber Bullying Detection on Social Media using Machine Learning." ITM Web of Conferences. Vol. 40. EDP Sciences, 2021.

[8] Talpur, KAZIM RAZA, SITI SOPHIAYATI Yuhaniz, and N. N. B. Amir. "Cyberbullying detection: Current trends and future directions." J. Theor. Appl. Inf. Technol. 98 (2020): 3197-3208.

[9] Shreenidhi B S, Mohammed Zaid Hulikatti, Nafey A H, Neha M R, Shradha."CYBERBULLYING DETECTION USING MACHINE LEARNING".

[10] Abutorab, Miss Jafri Sayeedaaliza, et al. "DETECTION OF CYBERBULLYING ON SOCIAL MEDIA USING MACHINE LEARNING".

[11] Nivethitha R Jayashree D Year: 2021" Cyberbullying Detection in Social Networks using Machine Learning Models" ICCAP EAI DOI: 10.4108/eai.7-12-2021.2314577.

[12] AGBAJE, MICHAEL, and Oreoluwa Afolabi. "Neural Network-Based Cyber-Bullying and Cyber-Aggression Detection Using Twitter Text." (2022).

[13] RAJESWARI, Mrs K., MUSHRUF BASHA M 2nd, and S. PRAVEEN. "Prevention and Suppression of Cyberbullying Using Machine Learning".

[14] Rao, Gauri Goyal, Mehul Wali, Diksha Yadav, Sarthak. (2022). "Cyber-Bullying Detection Using Machine Learning and Na¨ıve Bayes and N-Gram Model".

[15] Monirah Abdullah Al-Ajlan and Mourad Ykhlef, "Deep Learning Algorithm for Cyberbullying Detection" International Journal of Advanced Computer Science and Applications(IJACSA), 9(9), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090927.

[16] Moratanch N, Srimathi G, Sujitha S, Tharani M," CYBER BULLYING SCRUTINY FOR WOMEN SECURITY IN SOCIAL MEDIA"," irjmets,2023.

[17] Raj, C.; Agarwal, A.; Bharathy, G.; Narayan, B.; Prasad, M. Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques. Electronics 2021, 10, 2810. https://doi.org/10.3390/electronics10222810.