CYBERBULLYING DETECTION USING MACHINE LEARNING

[1]Dr. Sreenivas Mekala, [2]Y.Sai Nithin, [2]B.Guru Raghav, [2]T.Sai Kumar
[1]Associate Professor, Information technology, Sreenidhi institute of science and technology
[2]UG Scholar, Information technology, Sreenidhi institute of science and technology

Abstract: Cyberbullying occurs when someone harasses, threatens, or mistreats someone else online or on social media. Cyberbullying leads to threats, public embarrassment, and reprimands. Cyberbullying has caused a rise in youth suicide and mental health issues. It has lowered my self-esteem and escalated my suicidal thoughts. If cyberbullying continues, an entire generation of young adults will suffer from low self-esteem and mental health issues. Many machine learning algorithms are used to automatically identify social media cyberbullying. Social media monitoring allows this. These models don't account for all the factors that could make a comment or post bullying. Bullying is still ambiguous. This study presents a cyberbullying diagnosis model based on several factors. BERT, a bidirectional deep learning model, applies some of these factors.

Keywords: Cyberbullying, social media, BERT, NLP, semi supervised learning

I. INTRODUCTION

The term "social media" refers to online communities where people may share and discuss material such as images, videos, and documents. People use their laptops and mobile devices to communicate on social media. Social media platforms such as Facebook, Twitter, Instagram, TikTok, and many more enjoy widespread usage. These days, social media is used not only for nefarious purposes, but also in the realms of education and commerce. Moreover, social media is helping the global economy by generating a large number of new employment openings. There are many advantages to using social media, but there are also some disadvantages. Negative users engage in unethical and fraudulent behaviour on the internet in order to cause emotional distress and reputational harm to others. One of the most pressing problems with social media today is cyberbullying. The terms "cyberbullying" and "cyber-harassment" describe a form of bullying and harassment that takes place through the Internet. Online bullying can take many forms, including cyberbullying and cyber harassment. Cyberbullying, especially among adolescents, has become widespread with the expansion of the digital sphere and the development of related technologies. As a result of cyberbullying, victims and the people closest to them may sustain damage that cannot be repaired. Bullying, including cyberbullying, can result in a variety of negative outcomes, including low self-esteem, academic failure, fury, anxiety, despair, school avoidance, school violence, and even suicide.Cyberbullying can result in a variety of negative results, including damage to mental health, elevated levels of stress and anxiety, despair, aggressive behaviour, and low self-esteem.

It's possible that the emotional wounds caused by cyberbullying will remain long after the harassment has ended.An algorithm for detecting cyberbullying that is based on machine learning is something that we have proposed as a means of determining whether or not a given communication is related to cyberbullying. During the process of building the proposed model for the identification of cyberbullying, we have investigated a variety of different Machine Learning approaches. Some of these methods include Naive Bayes, Vector Machines for Support, Decision Tree, and Random Forest. Experiments are done utilising data gathered from posts and comments made on social media platforms such as Twitter and Facebook. The BoW and TF-IDF feature vectors are utilised in our performance evaluations. In terms of accuracy, the findings indicate that the TF-IDF feature is superior to BoW, however Random Forest exceeds any and all other machine learning methods. BoW is the most popular machine learning approach.

II. RELATED WORKS

Automatic cyberbullying identification using Twitter users' psychological characteristics was presented in 2020 by Vimala Balakrishnan et al. Twitter data collection, feature extraction, and cyberbullying detection and categorization are presented as the three primary steps towards better detection. There were a total of 9484 tweets in the annotated dataset, with 4.5 percent of people classified as bullies, 31.8 percent as spammers, 3.4 percent as aggressors, and 60.3 percent as typical. After the pre-processing step, during which non-English tweets, empty profiles, and special characters were removed, the final dataset had 5453 tweets. Textual, individual, and social network characteristics were the features extracted. WEKA 3.8 was used to run the model while 10-fold cross-validation was employed. Preliminary experimental study showed that Nave Bayes was not effective, therefore it was discarded in favour of Random Forest and J48. The classifiers were educated using data that was manually labelled.[1]

To produce contextual embeddings and task-specific embeddings, Google researchers Jaideep Yadav et al. developed a novel pretrained BERT model in 2020. A deep neural network known as the Transformer is used as the foundational model in the suggested approach. The Bert is constructed on top of a base model and uses its 12 layers to encode the incoming data. After being tokenized and appropriately padded, the input is sent into the model, which produces the final embeddings. In order to produce the final output, the embeddings produced by the previous layers must be classified by the classifier layer. They found that by using a pre-trained BERT model, they could improve upon the accuracy and consistency of earlier models for detecting cyberbullying. [2]

The method for identifying Twitter cyberbullying was proposed by Sudhanshu Baliram Chavan et al. in 2020.

To create the necessary dataset, we gathered information from places like GitHub and Kaggle. The TFDIF vectorizer algorithm is first applied to the data for preprocessing, during which features are extracted. The tweets are sent into a naive Bayes and support vector machine classifier for further analysis. When a tweet is identified as bullying, ten more tweets from the same account are retrieved and subjected to further analysis using naive Bayes and support vector machine classifiers. The tweet is deemed to be bullying if the user's chance of being bullied is more than 0.5. It was clear from the accuracy score and the outcomes that the SVM model was superior to the naive Bayes method. The SVM model achieved an accuracy score of 71.25%. [3]

John Hani et al. demonstrated a supervised learning method for identifying cyberbullying in 2019. During this phase, unwanted information and noise are eliminated from the data. Tokenization, word reduction, stop words, and encoding cleaning and word correction are used for this purpose. The second phase is feature extraction, and it makes use of TF IDF and sentiment analysis techniques like NGrams to take into account many permutations of words.

Kaggle's cyberbullying dataset is divided into a train and test set with a ratio of 0.8:0.2. Different n-gram language models are used to perform SVM and neural network classifiers. The effectiveness can be evaluated using metrics like accuracy, recall, precision, and f-score. It was discovered that Neural Networks outperformed SVM classifiers. The average f-score for a Neural Network was 91.9%, while the average f-score for an SVM was 89.8%. [4]

Using a convolutional neural architecture and modifying the concept of word embedding, the unique algorithm CNN-CB was proposed in 2018 by Monirah Abdullah Al-Ajlanet and Mourad Ykhlef. The structure is made up of four layers: the embedding, the convolution, the max pooling, and the dense layers. The convolutional layer, which compresses the input vector without losing significant features, receives the vector space of vocabulary created by the word embedding layer as its input. The Max pooling layer, the third one, takes the output of the second one as its input and calculates the maximum value of the selected region to preserve just the most important highlights. The identifying is performed by the final layer, called Dense.

The resulting accuracy was 95%. [5] Optimised Twitter cyberbullying detection based on deep learning (OCDD) was proposed in 2018 by Monirah A. Al-Ajlan et al. OCDD does not extract features from tweets; rather, it represents a tweet as a set of word vectors that are fed into a convolutional neural network (CNN) for classification. As a result, this method bypasses the need for selecting and extracting features. Word embedding, developed with the (GloVe) method, is used to represent interword semantic relationships. The many parameters in a CNN are optimised via a metaheuristic optimisation technique to yield the best possible classification results. CNN's ratings were quite high. [6]

Automatic cyberbullying detection using NLP, text mining, and ML was presented by Yee Jang Foong and Mourad Oussalah in 2017. The ASKfm dataset is based on user queries and profile excerpts from the social media platform ASKfm.

The pre-processing phase involves the elimination of hyperlinks and unfamiliar characters, the correction of any typos, and the replacement of lexicons with similar textual terms. The TF-IDF, the Unusual Capitalization Count, the Long-Iteration Word Count, and the Dependency Parser have all been used together. A 70% training set and a 30% test set are used. [7]

In 2016, X. Zhang et al. put forth an innovative strategy using a PCNN trained on pronunciation. Conversions from misspelt to correctly spelt words that share the same pronunciation are called "word-to-pronunciation" conversions. The starting point is determined by using two independent CNNs. We used word embedding from Google's word-vector as our first baseline feature collection. The second baseline, named CNN Random, relies on randomly generated vectors to construct its feature set. For PCNN's feature set, the phoneme codes were simply added as arbitrary vectors. Cost function adjustment was shown to be the most effective of the three strategies used to address class imbalance, which also included shifting the threshold and employing a hybrid approach. [8]

Michele Di Capua et al. developed a Growing Hierarchical SOMs-inspired design model in 2016 that might be used for cyberbullying detection without human supervision. Initially, characteristics are classified into four categories: syntactic, semantic, sentiment, and social. In order to properly categorise a sizable dataset, we employ the Growing Hierarchical Self Organising Map (GHSOM) network technique. It employs a multi-layered hierarchical structure, with different SOM types constituting each layer. The first layer just uses a single SOM. During this map's time period, a SOM could be added to the next level of the hierarchy for each unit. Using a K-fold partitioning of data, the GHSOM Network is trained and evaluated. [9]

Data mining and machine learning approaches were given as a method to identify cyberbullying by Sourabh Parime and Vaibhav Suri in 2014. Documents are clustered, data is pre-processed, an in-built classifier is used to generate labels from the features fed into it, occurrences are counted, a weight is assigned to each label, and irrelevant attributes are removed to help estimate the comment's nature during the text mining process, which is performed on unstructured data using machine learning techniques. The purpose of sentiment analysis is to identify the underlying mood of a piece of text. Data sets containing either positive or negative emotions are analysed separately. A supervised learning technique called SVM is trained using these, which are first stored in a vector. [10]

III. METHODOLOGY

To implement this method, we take advantage of the LDR's ability to change its resistance in response to changes in illumination. LDR is used to determine whether it is night or day in our suggested system. Then, the streetlight will be turned off during the day and back on at night automatically thanks to an infrared sensor that can identify the presence of a car on the road. If an infrared sensor detects a sparse vehicle population on the road, the lights will turn off. The lights will activate whenever a car is detected in the street. NLP (Natural Language Processing) refers to the first portion, and ML (machine learning) refers to the second; these are the two key components of the cyberbullying detection framework. In order for machine learning algorithms to work, they need some sort of formatting or structure added to the raw text. This necessitates a conversion of the algorithms to either vectors or integers before they can be used. This leads to the next phase, where the processed data is converted into a Bag-of-Words. Moreover, our model takes into account the TF-IDF metric. With the use of a statistical method called TF-IDF (phrase Frequency-Inverse Document Frequency), we may determine how important a given phrase is to a certain document within a larger corpus. We employ SMOTE and Xgboost for optimisation and improvement.

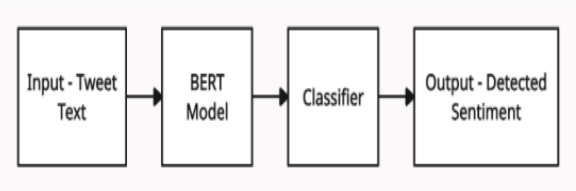

Fig.1. BERT model flow chart based on sentiment analysis

According to the restrictions established by its creator, the BERT model can only accept its input in a pre-processed form.

Better results have been obtained by the model thanks to these guidelines. All inputs are delivered to the model embedded as a mixture of the other three embeddings:

●Position embedding: BERT reads and uses existing embedding to express word order in a sentence. ●Segment Embedding: BERT can also take more than one sentence as input functions. It uses this embedding to understand the difference between two different sentences. ●Token Embeddings: This is the embedded text token from Word Piece token vocabulary.

The dataset represented here is a collection of tweets which was collected using Twitter API. The number of data entries exceeded 1000 tweets which belong to different time periods. The following images depict the datasets indicating Text Labels. Adaptation of the raw data according to our need is important before implementing the regression model. Since, raw data is most of the time inconsistent or incomplete or lacking in certain behaviour or lacking in attributes or may contain noises. So, we need to remove all these abnormalities and convert the dataset into something which can be used by the machine learning algorithms. So, we processed data obtained from online sources to obtain useful data metrics, related to profanity in the output, on a daily basis which can be used to train our models. The comment data which we downloaded was in xlsx format. So, we had to convert the xlsx file format to csv format which is usual format used to train machine learning models. Further sometimes data contain various inconsistencies such as noisy data which model cannot interpret and value dominances of a variable over another which can cause model's inconsistency to predict accurately

For training the model, first we import a specific algorithm class/module and create an instance of it. Then using that instance, we fit the model to the training data. Then we validate it by testing its accuracy score and fine tune its parameters till we get required results.

For testing the model, we compare its predicted values after the training phase with test data. Then input some different value for prediction and check whether it predicts it right. If it didn't predict right then, fine tune the algorithmic parameters and fit the model again.

Table 1: Test cases

TEST CASE	OUTPUT	STATUS
Comment1: u truly love nature	Identified as anti-bullying	Your comment is posted Successfully
Comment 2: Not a great picture Asshole	Identified as bullying	Warning! Your comment is filtered.
Comment 3: Beautiful pic	Identified as anti - bullying	Your comment is posted Successfully
Comment 4: Bullshit	Identified as bullying	Warning! Your comment is filtered
Comment 5: looks awesome	Identified as anti - bullying	Your comment is posted Successfully

IV. RESULTS AND DISCUSSION

A consequence of our testing based on the input processing and prediction that we carried out. For our model, we selected a tweet that contained traces of bullying. You can see the results of our categorization tests in Fig. 4. Here, the values 0 and 1 denote bullying and its absence, respectively. We used this data to create a confusion matrix, which is shown in Fig. 5. When applied to the same dataset as in [3], the accuracy of the SVM is 71.25 percent and the accuracy of the Naive Bayes model is 52.7 percent, as shown in Table 1. Using the BERT model for sentiment analysis on the Twitter dataset yields more precise results. When used to the Twitter dataset for sentimental analysis, our suggested model produced a higher accuracy of 91.90%, which can be deemed a better result when compared to the conventional machine

learning models employed on similar datasets.

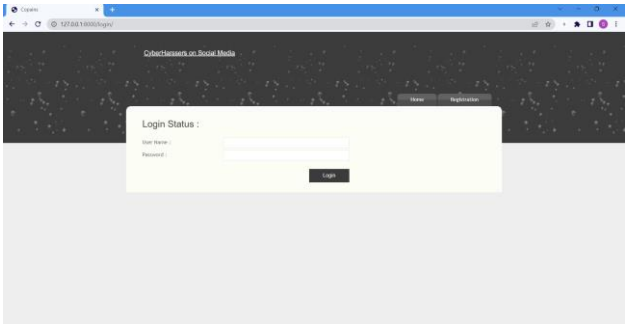

Fig. 2: This is the login page to the platform.

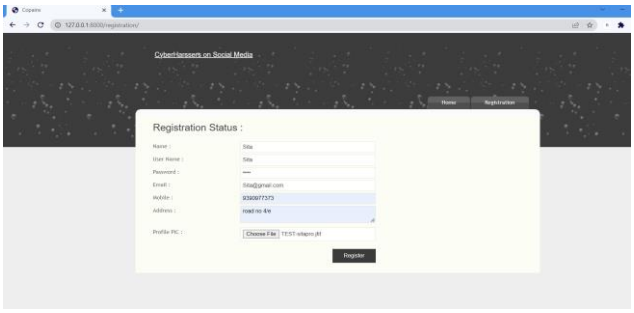

Fig. 3: It shows the new user registration process.

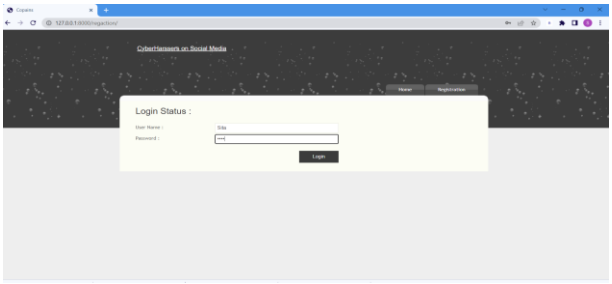

Fig. 4: This shows that one can login with their credentials.

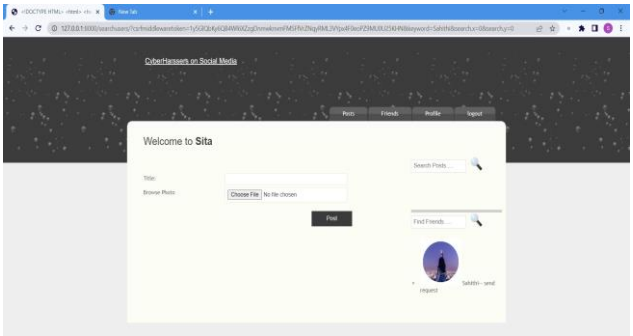

Fig. 5: This shows the users home page just after logging in.

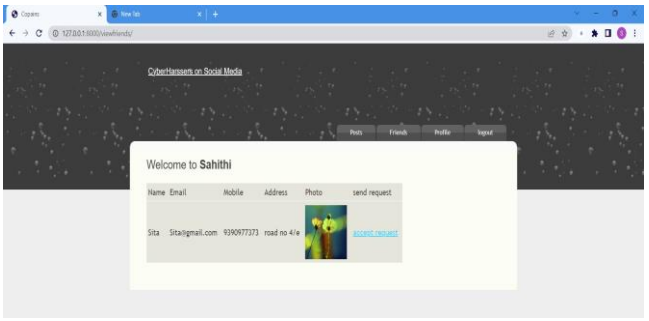

Fig. 6: This shows the other users web page where one can accept their friend requests.

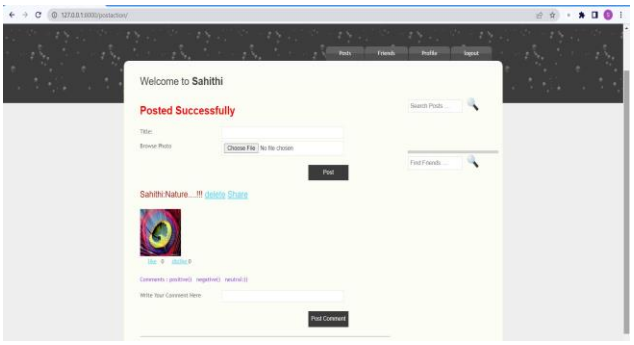

Fig. 7: This shows the post that a user can upload.

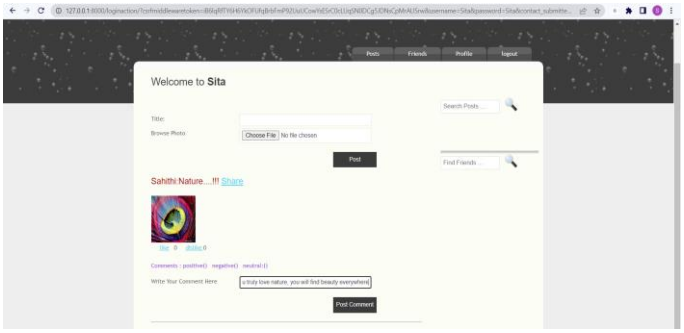

Fig. 8: This shows the anti-bullying comment that is posted by Sita on Sahithi's post.

Fig. 9: This shows that the anti-bullying comment is posted successfully.

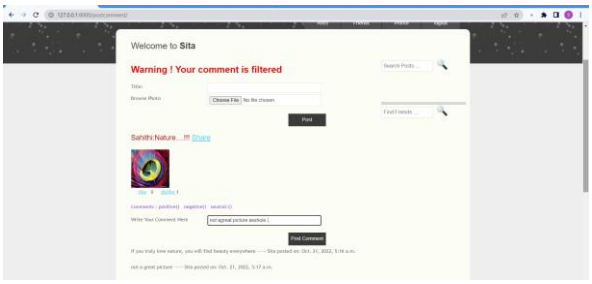

Fig. 10: This shows that the comment is filtered when a user tries to post a bullying comment.

V. CONCLUSION

Particularly, the proliferation of social media sites and the increased usage of social media by teenagers have contributed to the rise of cyberbullying and other serious societal challenges. Bad outcomes from cyber harassment can be avoided if a system for automatically detecting cyberbullying is developed. Taking into account two features, BoW and TF-IDF, we explored the automated identification of social media posts connected to cyberbullying in this study, which highlights the importance of detecting cyberbullying. In this study, we apply four different machine learning methods to detect bullying text, including SVM for both BoW and TF-IDF. In the future, we hope to use deep learning algorithms to automatically detect and categorise cyberbullying from Bengali literature within a unified framework.

REFERENCE

[1] Balakrishnan, Vimala & Khan, Shahzaib & Arabnia, Hamid. (2020). Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning. Computers & Security. 90. 101710. 10.1016/j.cose.2019.101710.

[2] J. Yadav, D. Kumar and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 1096-1100, doi: 10.1109/ICESC48915.2020.9155700.

[3] R. R. Dalvi, S. Baliram Chavan and A. Halbe,"Detecting A Twitter Cyberbullying Using Machine Learning," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893

[4] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer and Ammar Mohammed, "Social Media Cyberbullying Detection using Machine Learning" International Journal of Advanced Computer Science and Applications (IJACSA),10(5),2019.

[5] 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

[6] M. A. Al-Ajlan and M. Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning," 2018 21st Saudi Computer Society National Computer Conference (NCC), Riyadh, 2018, pp. 1-5, doi: 10.1109/NCG.2018.8593146.

[7] Y. J. Foong and M. Oussalah, "Cyberbullying System Detection and Analysis," 2017 European Intelligence and Security Informatics Conference (EISIC), Athens, 2017, pp. 40-46, doi: 10.1109/EISIC.2017.43.

[8] X. Zhang et al., "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA,2016, pp. 740-745, doi: 10.1109/ICMLA.2016.0132.

[9] M. Di Capua, E. Di Nardo and A. Petrosino, "Unsupervised cyber bullying detection in social networks,"2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, 2016, pp. 432-437, doi: 10.1109/ICPR.2016.7899672.

[10] S. Parime and V. Suri, "Cyberbullying detection and prevention: Data mining and psychological perspective," 2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT- 2014], Nagercoil, 2014, pp. 1541-1547, doi: 10.1109/ICCPCT.2014.7054943.