Journal of Harbin Engineering University Vol 44 No. 8
ISSN: 1006-7043 August 2023

Cyber Bullying Detection Using Artificial Intelligence

[1] Alima Hassan, [2] Prof. Vanita Tank
[1] Student, Dr. Vishwanath Karad MIT World Peace University, Pune, India
[2] Professor, Dr. Vishwanath Karad MIT World Peace University, Pune, India

Abstract: Social media has given us a lot of opportunities and benefits. In spite of all the benefits, people are still getting bullied by anonymous users. When someone bullies another person using technology, they may send or post offensive or hateful messages about them. This is known as cyberbullying. With the extensive use of social media users, cyber bullying has also increased. Hence, it's very important to detect cyber bullying on online platforms. This enables large-scale social media monitoring. Here we have implemented this project to identify the hate speech or offensive speech from twitter. Main aim of our project is to detect cyber bullying in tweets using Machine learning and deep learning classification algorithms like Logistic Regression, Naive Bayes, LSTM (Long Short-Term Memory) and CNN (Convolution Neural Network). NLTK (Natural Language toolkit) is also used to preprocess text data. Then compare all the algorithms used, check the accuracy of each model and then choose the best model for detecting cyber bullying in tweets. The motivation behind this project is to protect our society from cyber bullying and also to prevent youngsters and teenagers from getting bullied, committing suicide because of bullying and reducing crime in cyberspace.
Index Terms: Cyber bullying, Twitter, Machine Learning, Deep Learning, Natural Language Toolkit.

Ⅰ. Introduction

Cyberbullying is the practice of hurting another person's feelings, emotions, or thoughts by using derogatory language in messages or posts sent or posted on websites or online services. It is done to intentionally hurt others emotionally, mentally and physically. In recent years, social media has taken over as the main channel for distributing ideas around the globe. People can now stay in touch with one another despite cultural and economic barriers, thanks to the rise in the use of social networking sites. Although social media is safe to use, because so much information is shared there, it is challenging to censor texts, which encourages hate speech. In this study, we presented our strategy for addressing hate speech and, to a large extent, reducing it.

In recent years, hate speech has dramatically increased. In reality, because all work and communication has since been conducted online, the situation has gotten worse since the COVID-19 pandemic-related shutdown. Because more people are using social media, there are more instances of cyberbullying. These platforms offer a free platform for users to express their opinions, share, orconvey and messages to people all over the world, but as a lot of content is being shared on these platforms so it's not possible to control the content. To address abuse, cyberbullying, illegal activity, sexual assault, and violence against public figures, Facebook has established a set of community standards. Twitter also has some rules that can help someone who has been the victim of social abuse, similar to Facebook.

Cyberbullying is a more significant issue than traditional bullying, according to research from the University of British Columbia. According to 733 surveys of adolescents, 12 percent of them had engaged in traditional bullying, compared to 25-30% who had engaged in cyberbullying. 95 percent of them claimed that they only make fun of people online by mocking them, while the other 5 percent admitted that their original intent was to hurt or insult them. Teenagers significantly underestimate the risks of cyberbullying, according to the report. So, our focus is to create a model to identify cyberbullying. Tweets are typically divided into bullying and non-bullying tweets using text classification based on supervised machine learning (ML) and deep learning models. Dealing with hate speech on Twitter is our main focus. There are charts and tables that show the degree of accuracy attained by various models.

Ⅱ. Literature survey

Cyberbullying detection is gaining importance and is receiving a lot of attention online. Finding ways to identify cyberbullying on social media has grown to be a crucial research area. In the past few years, cyberbullying has contributed to suicides, and India is one of the four nations with the highest number of victims. Due to an increase in cases since 2015, it is now required in universities and schools to prevent cyberbullying. A recent study uses machine learning and deep learning techniques to automatically detect comments that are cyberbullying. Accuracy, precision, recall and F1-score, are employed to assess the performance of the model. In their research, a deep learning technique called Gated Recurrent Unit outperformed all others with an accuracy of 95.47% [1]. In another research [2], they propose a system to give a double characterization of cyberbullying to monitor bullying and harassment in virtual environments and executed it with the help of machine learning and language preprocessing. Their method makes use of CNN's creative idea for analysing the content. Their system achieves precise results by employing an accurate method of CNN. Another similar study [3] looked into the automated detection of social media posts relating to cyberbullying by taking into account the two features BoW and TF-IDF. Another study examined the literature research on different machine learning algorithms and found that the Naive Bayes N-gram provided the highest level of accuracy [4]. A new technique for identifying hate speech on Twitter uses unigrams, sentimental features, and semantic features to automatically distinguish between bullying and non-bullying tweets. Their methods were 87.4 percent accurate in classifying tweets as binarily offensive or non-aggressive, and 78.4 percent accurate in ternary form as hateful, separate, violent, or naive. [5]. Machine learning is used in a similar study [6] to address the issue of cyberbullying on the Twitter platform. The experiments made use of both supervised and unsupervised machine learning techniques. It was found that choosing the appropriate set of keywords is crucial for improving sentiment analysis outcomes, particularly when performing topic modelling. On the dataset retrieved using particular keywords, the K-means and LDA algorithms both produced poor results. Any classification task's outcomes can be correlated with the quality of the annotated corpus. When used on an annotated corpus instead of a manually labelled dataset, SVM classifier performed better. Thus, it was concluded that the annotator's comprehension of the tweet plays a major role in labelling. Table 1. shows the literature survey conducted in order to fill gaps in previous studies.

Table 1. Literature review

Sr. no.	Author	Paper Name	Method	Result
1.	Apoorva K G; D Uma	Detection of Cyber bullying Using Machine Learning and Deep Learning Algorithms	Used Machine learning and deep learning algorithm.	Obtained highest accuracy with Gated Recurrent, a deep learning technique unit algorithm
2.	Saloni Mahesh Kargutkar Prof. Vidya Chitre	A Study of Cyber bullying Detection Using Machine Learning Techniques	CNN is used with multiple layers and offers an iterative analysis process over various layers to provide an accurate and efficient analysis.	CNN Implementation using keras. It helps in achieving precise and accurate results.

3.	Manowarul Islam, Linta Islam, MD Ashraf Uddin, Arnisha Akhter	Cyber bullying Detection on social networks using Machine learning Approaches	Automated identification of posts on various social media sites related to cyber bullying using two features BoW and TF-IDF.	SVM better than other algorithms. TF IDF better performance and accuracy than BoW.
4.	Muskan Patidar, Mahak Lathi, Manali Jain, Monika Dhakad, Prof. Yamini Barge	Cyber Bullying Detection for Twitter Using ML Classification Algorithms	Naïve Bayes unigram, Naïve bayes bigram, Naïve Bayes trigam, Naïve Bayes N- gram	Naïve Baiyes provides the best accuracy. Identify bullying and non-bullying using ML algorithms
5.	M. Bouazizi, H. Watanabe and T. Ohtsuki	"Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection"	Patterns and unigrams are used to train a ML algorithm	Performed Hate speech detection On twitter using ML
6.	A. Shekhar and M. Venkatesan	"A Bag-of-Phonetic-Codes Model for Cyber-Bullying Detection in Twitter"	Bag-of-Phonetic-Codes model for extracting textual features	Detecting cyber bullying using a Novel method
7.	M. Gomes, R. Martins, J. J. Almeida, P. Henriques and P. Novais	Hate Speech Classification in social media Using Emotional Analysis	Textual features, primary datasets, and machine learning models are all used.	Analysis of several hate speech datasets
8.	S. S. Syam, B. Irawan and C. Setianingsih	Hate Speech Detection on Twitter Using Long Short-Term Memory (LSTM) Method	LSTM method as a classifier	Identify a tweet that is solely based on a hashtag search on Twitter.
9.	Mohammed, Nora	Extracting Word Synonyms from Text using Neural Approaches. The International Arab Journal of Information Technology	Novel 2-step approach, Natural Language Processing (NLP)	Train a neutral network to distinguish between synonyms and other words with similar meanings using an annotated dataset.

Ⅲ. System architecture

The issue of cyberbullying detection on the Twitter platform is the main focus of this study. The primary tasks in combating cyberbullying threats are the detection of cyberbullying activities from tweets and the necessary preventive measures. This is because cyberbullying has grown to be a significant issue on Twitter. Therefore, it's crucial to conduct research on cyberbullying and develop new tools and technologies to address the problem. It is impossible to manually stop cyberbullying on Twitter. Additionally, it is very challenging to detect cyberbullying by mining social media messages. It is impossible to infer intentions and meanings from tweets because they frequently use informal language. Additionally, it is

more difficult to spot the bully if they use sarcastic techniques. The detection of cyberbullying is a significant and active research area because of all the difficulties that social media messages pose. On twitter platform, the detection of cyberbullying has been pursued using machine learning, deep learning, and tweet classification methods.

Tweets are divided into bullying and non-bullying tweets using text classification based on supervised machine learning (ML) models and deep learning models. When the class labels are fixed and irrelevant to the most recent events, supervised classifiers perform poorly. In order to form the patterns or classes in the entire dataset, important topics from a set of data are traditionally extracted using topic modelling approaches. Although the concept is the same, short texts cannot be effectively analysed using general unsupervised topic models; as a result, special unsupervised short text topic models were employed. These models are efficient at extracting the trending topics from tweets for further processing. Utilizing the bidirectional processing to extract significant topics is made easier with the aid of these models. To obtain sufficient prior knowledge, which is not always enough for these unsupervised models, extensive training is necessary. A strong method should be used for classification of tweets which can fill the gap between classifier and model. In this article, we suggest a method based on machine learning and deep learning to detect bullying in tweets.

We first fitted the model using data from 10 epochs, and as a result, we got an accuracy of about 94%.

Epochs are the number of times the training set is exposed to the model. The inside model parameters may be refreshed by the preparation dataset. At least one clump is present in an epoch. Then, to determine whether it is hate speech or not, we suggest using Natural Language Processing, Logistic Regression, Naive Bayes, LSTM, and CNN.

Advantages:

- Highest accuracy
- Reduces time complexity.
- Easy to use

Our system's architecture is based on client server architecture, in which clients make service requests and the server responds to them. HTML, CSS, and JavaScript are used on the server side, and Python, Flask, MySQL, machine learning, and deep learning models are used on the client side. The server responds to the client's request for data by accommodating it and returning the requested data packets to the user. Flask is used for connectivity of front end and back end. To access database data, SQL Yog is used. We can connect to the SQL Yog database with the aid of Xampp Controller.

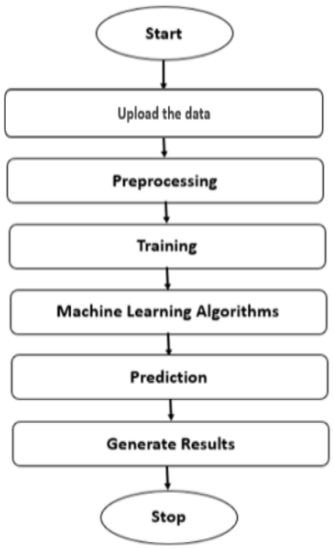

Fig 1. Block Diagram

The model as shown in fig 1. includes the following steps: (i) Data collection, (ii) Pre-processing and data cleansing, (iii) Feature extraction and feature selection

(iv) Classification.

A. Data collection:

The input dataset is made up of tweets that were streamed by the Twitter API and contained a total of about 32 cyberbullying keywords. Some of the terms suggested by psychology literature include idiot, shit, donkey, live, afraid, moron, poser, rape, ugly, bitch, whale, and LGBTQ (le***an, g*y, bisexual, transgender, and queer). While other terms like "ban," "kill," "die," "evil," "hate," "attack," "terrorist," "threat," and "racism," were proposed. The initial dataset consists of 435764 records, with 130000 tweets based on keywords

related to racism, insults, swear words, and sexism. There are a lot of outlier tweets in this dataset. Since the tweets must be in English, any messages containing terms from other languages are deleted, and retweets are filtered. After removing those kinds of irrelevant tweets, randomly select about 10,000 tweets from the remaining ones to create a complete dataset as shown in fig. 2. As part of the pre-processing stage, each of these actions is performed automatically. The final pre-processing steps are carried out after that.

	id	text	label	type	model_wr	db.model	status	round	split	annotator
2	1	10000 It should t	hate	notgiven	TRUE	0.97457	dynabench	1	train	gnzuctwed3
3	2	10001 dalits are l	hate	notgiven	FALSE	0.08233	dynabench	1	test	cagnluiznm
4	3	10002 dalits are l	hate	notgiven	TRUE	0.92319	dynabench	1	train	cagnluiznm
5	4	10003 It was a bi	hate	notgiven	TRUE	0.99006	dynabench	1	test	gnzuctwed3
6	5	10004 I don't wo	hate	notgiven	TRUE	0.98836	dynabench	1	train	cagnluiznm
7	6	10005 I don't wo	hate	notgiven	TRUE	0.99506	dynabench	1	train	cagnluiznm
8	7	10006 I don't wo	hate	notgiven	TRUE	0.9934	dynabench	1	train	cagnluiznm
9	8	10007 I don't wo	hate	notgiven	TRUE	0.98625	dynabench	1	train	cagnluiznm
10	9	10008 I don't wo	hate	notgiven	TRUE	0.95252	dynabench	1	test	cagnluiznm
11	10	10009 I don't wo	hate	notgiven	FALSE	0.09288	dynabench	1	train	cagnluiznm
12	11	10010 I don't wo	hate	notgiven	FALSE	0.46144	dynabench	1	train	cagnluiznm
13	12	10012 foreigners	hate	notgiven	TRUE	0.98753	dynabench	1	train	cagnluiznm
14	13	10013 immigrant	hate	notgiven	TRUE	0.98971	dynabench	1	train	cagnluiznm
15	14	10014 women ar	hate	notgiven	TRUE	0.9814	dynabench	1	dev	cagnluiznm
16	15	10015 gay peopl	hate	notgiven	TRUE	0.53936	dynabench	1	train	cagnluiznm
17	16	10016 gay peopl	hate	notgiven	TRUE	0.8682	dynabench	1	train	cagnluiznm
18	17	10017 Why is it t	hate	notgiven	TRUE	0.89113	dynabench	1	train	cagnluiznm
19	18	10018 Why is it t	hate	notgiven	TRUE	0.96767	dynabench	1	train	cagnluiznm
20	19	10019 Why is it t	hate	notgiven	TRUE	0.92169	dynabench	1	train	cagnluiznm
21	20	10020 Why is it t	hate	notgiven	TRUE	0.92405	dynabench	1	train	cagnluiznm
22	21	10021 Why is it t	hate	notgiven	TRUE	0.83432	dynabench	1	train	cagnluiznm
23	22	10022 Why is it t	hate	notgiven	FALSE	0.37971	dynabench	1	train	cagnluiznm
24	23	10023 Why is it t	hate	notgiven	TRUE	0.56723	dynabench	1	dev	cagnluiznm

Fig 2. Twitter Dataset Obtained from Kaggle.com

B. pre-processing and data cleansing

The phases of preprocessing and data cleaning have three subperiods. This process will be applied to the raw tweet dataset to generate a final result. The first subphase involves noise removal procedures like URL removal, hashtag removal, keyword removal, punctuation removal, emoticon removal, and transformations. The second sub-phase involves the removal of repeated characters from sentences and other vocabulary cleaning tasks like spell checking, acronym expansion, slang modification, and elongation. The last sub-phase involves the transformation of tweets using techniques like stemming, tokenization and stop-word filtering. These sub sections must be completed to increase tweet quality, feature extraction, and classification accuracy. Cleaning up the data is the process of data preprocessing. Every process has this crucial step, which needs to be handled with extreme care. Raw data must be transformed into the necessary form of data in order for the model to be trained properly. For instance, the data in the raw form is "You look so fat and ugly, change your style," while the data after pre-processing is "look fat and ugly, change your style." The pre-processed data eliminates all the unnecessary words and special characters that are not needed for model training, such as what, who, with, is, the, etc. Data is split down into sentences, and each sentence is then padded with a common word to make sure it has an even number of words. This aids in ensuring the uniformity of the data. Additionally, data is converted into lowercase format and then into a vector before being sent to the model in that format.

C. Feature extraction and selection

Using NLP tools like Word2Vec and TF-IDF, the features from the dataset are extracted. The verbs and adverbs here serve as supplemental details to the nouns, pronouns, and adjectives, which serve as the main feature contents. Additionally, the classification performance can be enhanced by extracting Part-of-Speech (POS) tags, function words, and other features. There are numerous techniques for choosing features. Prominent features are chosen in order to identify the cyberbullying events using a variety of techniques, and these feature subsets are then fed into various ML and deep learning techniques.

D. Classification

The texts or comments from twitter dataset are classified into two types as follows:

- Non-bullying Text: These are not bullying remarks or supportive remarks; rather, they are comments. These are positive and non-bullying comments. "This photo is very beautiful," for instance.
- Bullying Text: This type of text belongs to bullying and harassment. For instance, the text or comment "go away bitch" is considered a negative comment.

Ⅳ. Methodology

- In this project, we have used Python, Machine learning, deep learning and web technology as shown in fig 3.
- First search for the twitter dataset from

Kaggle.com and download it for training purpose of the model.

- Following column extraction from the downloaded dataset, the data is cleaned by removing all non-alphabetic and special characters (such as #, @), stop words, and stop phrases using the NLTK stop word corpus. The next step is to tokenize the data by writing a collection of hate speech and cyberbullying messages.
- Train the model with the help of Data set.
- Data splitting is done. 70% for training and 30% for testing purposes is used.
- Apply the generated model to the fetched tweets and then user is supposed to write any message on the frontend and then propose Natural Language Processing, Logistic Regression techniques, Naïve Baiyes, LSTM and CNN to predict whether it is cyber bullying or not.
- Then comparing different algorithms, checking the accuracy of each model and then choosing the best model.
- After checking the accuracy, got the highest accuracy for logistic regression as it is a great machine learning algorithm for text related classification and used when the data is in the form of binary i.e.; 0 and 1. It is the best algorithm used to solve binary classification problems. So, will use logistic regression to analyse the input given by user.
- For the frontend, create a platform where the user can login and post a tweet and then system uses the machine learning model (as logistic regression got highest accuracy) to analyse the input given by user whether it is a cyber bullying tweet or not.

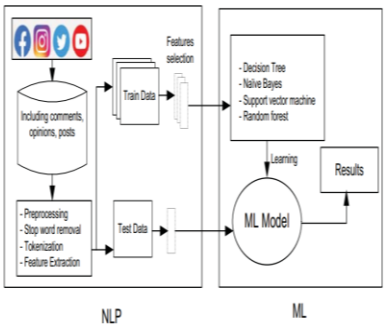

Fig 3. Proposed Model

A. **Natural Language Processing**

In real-world posts and texts, there are a lot of informal language used which includes characters and words. We must prepare the machine learning algorithms for the upcoming stage before using them. Tokenization, stemming, and the elimination of any extraneous characters, such as stop words, punctuation, and numbers, are among the processing steps in this phase. Following the preprocessing, we prepare the texts' two most crucial features as follows:

1) Bag-of-Word: Algorithms for machine learning are incompatible with the raw text. Therefore, we must convert the algorithms to vectors or numbers before applying them. It can distort the text's semantic meaning and doesn't remember any original textual order. The order of the words doesn't really matter; it's just counting how many times each one appears in a passage. Every word count turns into a dimension for that particular word. Therefore, for this stage, the data is transformed into Bag-of-Words (BoW).

2) TF-IDF: This feature is also taken into account by our model. Term Frequency-Inverse Document Frequency, a statistical tool, can determine how relevant a word is to a document within a collection of documents. It determines the "normalized count," which divides each word count by the number of documents in which a given word appears. Every word in a bag of words has equal importance, and in TFIDF, the more common words should also be given more weight because they are helpful for categorization.

B. Hardware and software requirements

Hardware Configuration:

- Operating system: Windows 7 or 7+
- RAM: 4 or 8 GB
- Hard disc or SSD: More than 500 GB
- Processor: Intel 3rd generation or high or Ryzen with 4 or 8 GB Ram

Software Configuration:

- Software's: Python 3.6 or high version
- IDE: PyCharm.
- Framework: Flask
- Database: SQLyog
- XAMPP Control Panel

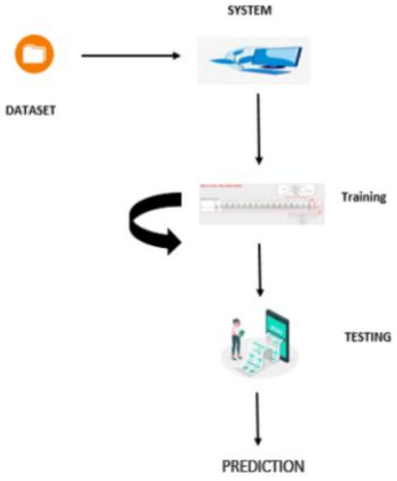

Fig 4. Architecture of the system

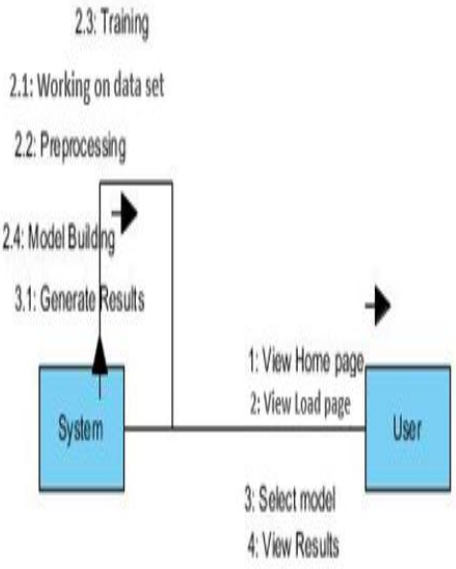

Fig 5. Collaboration Diagram

Ⅴ. Results

The system can accurately determine whether the entered text constitutes cyberbullying or not. We have achieved the following output.

Fig 6. Home Page

Fig. 7 displays the Home page that we developed using HTML, CSS, and JavaScript on the front end.

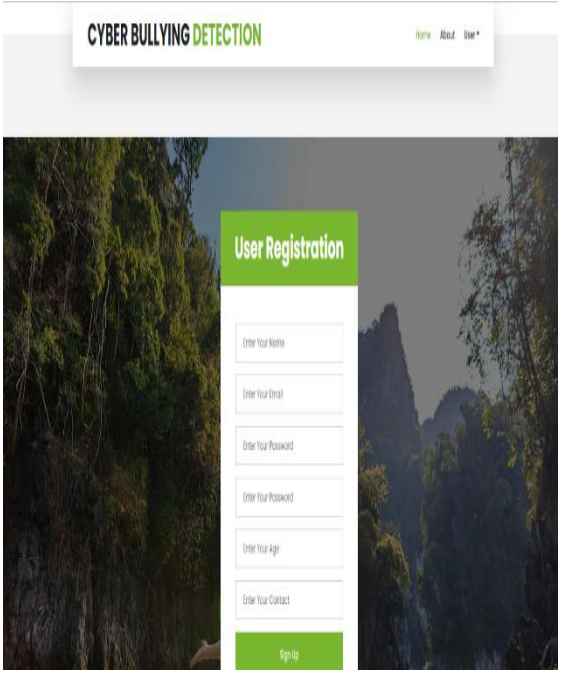

Fig 7. User Registration

Fig. 8 shows the user registration page where user can register

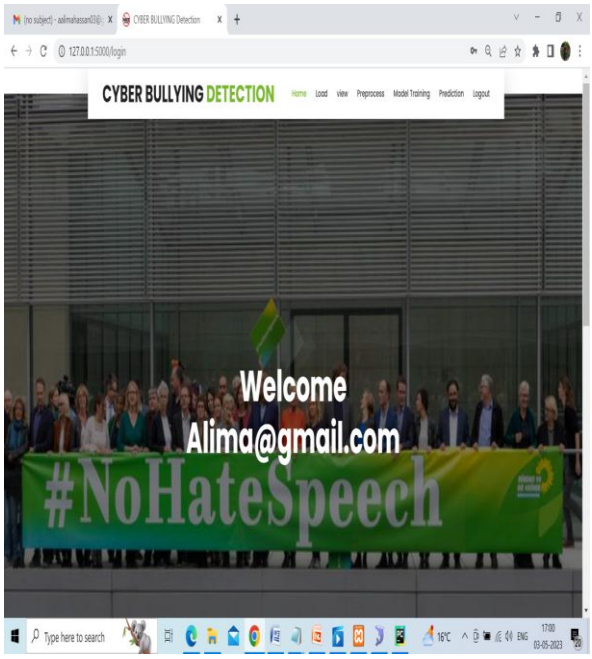

Fig 8. Login Page

Fig 8. Shows the login page where we can login with the help of credentials i.e.; username and password. After entering the login credentials, we will be directed to load page where we have to upload the dataset.

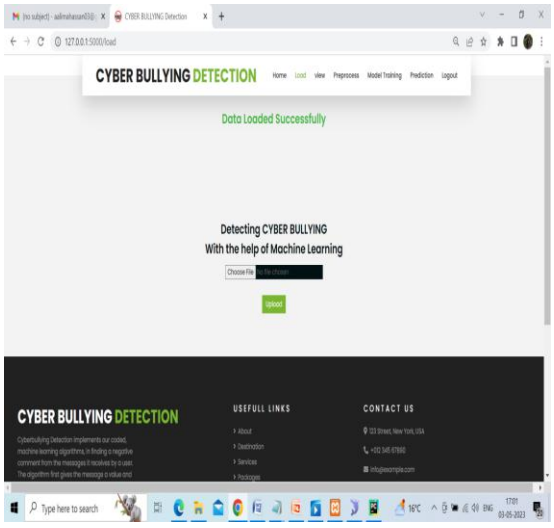

Fig 9. Load Page where we load the dataset

Fig 9. Shows the load page where we can upload the dataset which is in the CSV format.

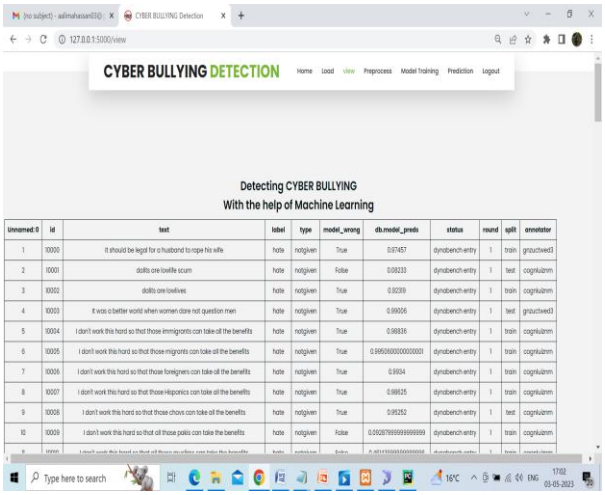

Fig 10. View the loaded dataset

Fig 10. Shows the view page where we can view the loaded dataset.

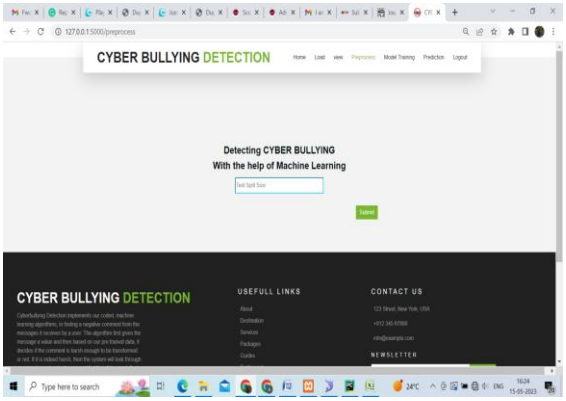

Fig 11. Preprocess

Figure 11 illustrates the preprocess page where the dataset can be divided into training and testing portions. Here, the training set is used to train the model, and the testing set is used to test it. To accurately predict a particular outcome that the model wants to predict, a machine learning algorithm or model is trained. Here we have used 70% data for training and 30% for testing purpose because we got the highest accuracy by keeping it 70% and 30%.

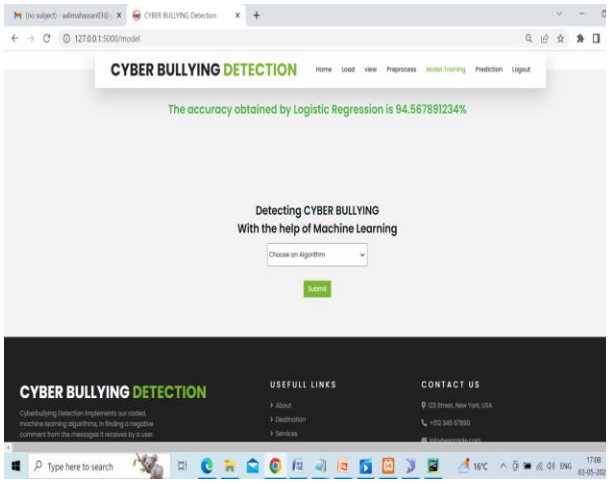

Fig 12. Model Training Page

Here we train the model with the help of different Machine learning and Deep learning algorithms and choose the best algorithm with highest accuracy. As shown in Fig 12. Logistic regression got the highest accuracy i.e.; 94.56%

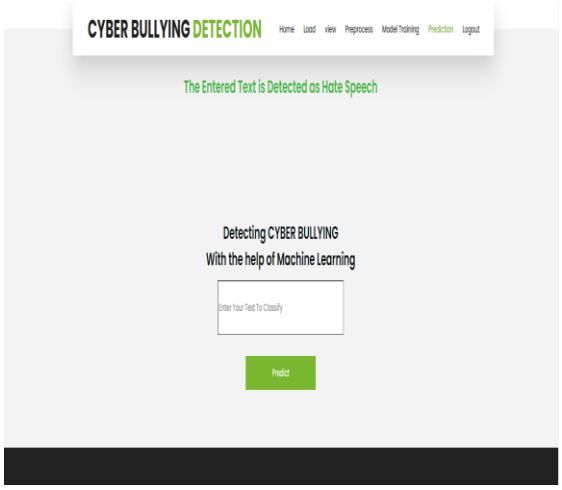

Fig 13. Prediction Model

The user enters text here, after which keywords are found in the sentence and their frequency in the pickle files is checked from the bag of words. If the frequency value of the keyword exceeds the threshold value, the result is generated. The result is either whether the content is cyber bullying text or not.

Accuracy

The percentage of accurate predictions made by the model is referred to as accuracy. Below is a graph of the accuracy of different machine learning algorithms. We contrast the various machine learning algorithmic parameters using the two significant feature vectors, BoW and TF-IDF. We looked at these findings and discovered that TF-IDF performs more accurately than BoW. Logistic regression performs better than the other machine learning algorithms.

Accuracy = (True Positive + True Negative) / Total Samples (1)

In this section, the proposed work's accuracy is contrasted with that of earlier works.

Table 2. Accuracy Table

Model	Accuracy obtained
Logistic Regression	94%
Naïve Bayes	65%
LSTM	78%
CNN	68%

Table 3. shows accuracy obtained by training 70% data and testing 30% data in the proposed system. Based on how well they work with the dataset, various machine learning and deep learning algorithms are used in the proposed work. The highest accuracy, 94%, is provided by logistic regression, which is significantly higher than any accuracy found in any of the earlier works. Logistic regression, one of the simplest algorithms, is easy to use and, in some cases, provides excellent training efficiency. To determine a single dependent variable with only two possible outcomes, such as pass/fail or yes/no, logistic regression is used. Similar to classification, it is most effective in circumstances with binary outcomes. The model may be dependent on one or more independent variables. Predicting the likelihood of an event is useful when using logistic regression analysis. It aids in calculating the odds between any two classes. In a nutshell, logistic regression can determine whether a post or comment is cyberbullying text by analysing the data. Small datasets are well suited for naive bayes. The accuracy achieved is noticeably greater than any of the earlier works completed. This is a result of the data preprocessing techniques used in

the suggested work and the use of a sizable portion of the data for training. Due to the textual nature of the data set,
other algorithms produce results with slightly lower accuracy. The tweets frequently use informal language, which makes it difficult to obtain accurate results. Additionally, as the data set changes, the accuracy will change as well.

Table 4. Accuracy table obtained in similar Work [3]

Model	Accuracy obtained in previous similar work [3]
Decision Tree	78%
Naïve Bayes	76%
Support vector Machine	75%
Random Forest	75%

The accuracy for various Machine Learning algorithms obtained in the related literature review work is shown in Table 4 [3]. Implementing a decision tree classifier yields the best results, but because of a different data preprocessing technique used and difference in training and testing size of the dataset, the accuracy is still lower than that of the proposed work.
Python machine learning packages are used to implement the algorithms for detecting bullying. The following metrics are used to analyse the performances. The confusion matrix, also known as the contingency table, contains a list of the categorization results.
The True Positive box in the upper left corner displays the number of people who were reported as true positives when they were true. The number of samples that were incorrectly categorised as false negatives is reflected in the False-positive bottom right cell. False-negative counts the number of people who were counted as true even if they were false. False-positives reflect the amount of people who were listed as true because they were actually true.

Table 5. Confusion Matrix

	Condition Positive	Condition Negative
Predicted condition positive	True Positive	False Negative
Predicted condition negative	False Positive	True Negative

$$\text{Precision} = \frac{\Sigma \text{ True Positive}}{\Sigma \text{ Predicted Condition Positive}} \quad (2)$$

$$\text{Recall} = \frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}} \quad (3)$$

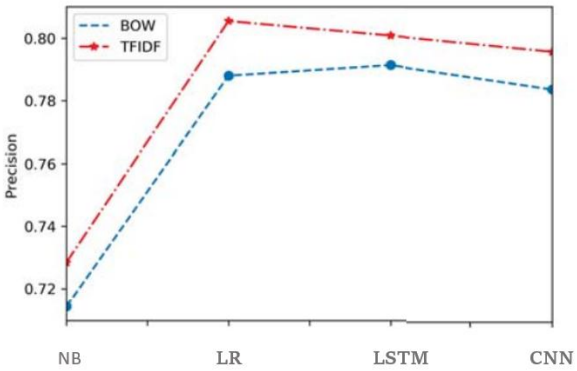

Fig 14. Accuracy of dataset based on confusion matrix

A graph representing the precision and accuracy of different machine learning methods is displayed. We got the similar outcomes and discovered that TF-IDF performs more accurately than BoW. The most effective algorithm is Logistic Regression, followed by the others.

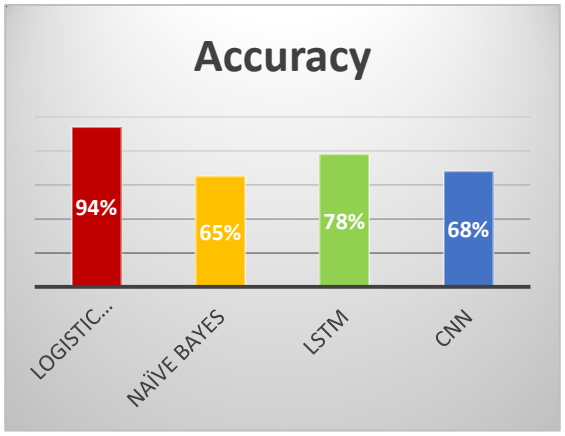

Fig 15. Accuracy of Dataset

Ⅵ. Conclusion

With the growing popularity of social media sites and the rise in teen social media use, in particular, cyberbullying has increased in frequency and has started to cause serious social problems. To prevent negative effects of online harassment, an automatic method for detecting cyberbullying must be developed. We investigated how to automatically recognize posts on social media that were associated with cyberbullying by taking into account two features: TF-IDF and BoW. Using text classification algorithms, we attempted to locate cyberbullying in the Twitter data. Although we have used a variety of machine learning and deep learning algorithms for text-based classification, including Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and others, logistic regression performed better than all of them. Given that this research has a wide range of sub-problems, its future direction is always up for debate. We also compared our work with another similar work and found out our Logistic regression outperformed their classifiers as we got the highest accuracy. Hence, our work is definitely going to help society from the prevention of cyberbullying so that they can use social media safely. However, cyberbullying detection is limited by the size of training data. As twitter dataset contains a lot of informal language so it was quite difficult to achieve good accuracy for deep learning techniques but in future, if a larger dataset is used and preprocessing methods are done carefully then deep learning techniques will be suitable for larger dataset. To find more posts on social media that are related to cyberbullying, the textual component can be taken into account alongside the image. Additionally, we can use additional deep learning network models to improve prediction accuracy of larger datasets.

References

[1] A. K. G and D. Uma, "Detection of Cyberbullying Using Machine Learning and Deep Learning Algorithms," *IEEE 2nd Asian Conference on Innovation in Technology (ASIANCON),* Ravet, India, pp.1-7, 2022.

[2] S. M. Kargutkar and V. Chitre, "A Study of Cyberbullying Detection Using Machine Learning Techniques," *Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, pp. 734-739, *2020*.

[3] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin and U. K. Acharjee, "Cyberbullying Detection on Social Networks Using Machine Learning Approaches," *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Gold Coast, Australia, pp. 1-6, 2020.

[5] M. Bouazizi, H. Watanabe and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," in *IEEE* Access, vol. 6, pp. 13825-13835, 2018.

[6] A. Shekhar and M. Venkatesan, "A Bag-of-Phonetic-Codes Model for Cyber-Bullying Detection in Twitter," *International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, India, pp. 1-7, 2018.

[7] S. S. Syam, B. Irawan and C. Setianingsih, "Hate Speech Detection on Twitter Using Long Short-Term Memory (LSTM) Method," *4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE),* 2019, pp. 305-310, 2019.

[8] M. Gomes, R. Martins, J. J. Almeida, P. Henriques and P. Novais, "Hate Speech Classification in Social Media using Emotional Analysis," *7th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 61-66, 2018.

[9] Mohammed, Nora, “Extracting Word Synonyms from Text using Neural Approaches,” *The International Arab Journal of Information Technology*, pp. 45-51, 2018.