ISSN 2277-2685
IJESR/Sep. 2023/ Vol-13/Issue-3/1-09
Mrs.Y.Prathima *et. al.*, / International Journal of Engineering & Science Research

Detection of Cyberbullying on Social Media Using Machine learning

Mrs.Y.Prathima[1], Saranya Chakravarthy Korrapati[2], K Nikhil Chary[2], Chiliveri Rishkeswa[2], Shaik Kaif[2]*[1]Asst. Professor,Computer Science and Engineering, CMR Engineering College, medchal, T.S, India[2]B.Tech, Computer Science and Engineering, CMR Engineering College, medchal, T.S, India*

Abstract: Cyberbullying is a significant issue on the internet that affects both adults and adolescents.Mistakes like despair and suicide have resulted from it. A increasing demand exists for the regulation of material on social media platforms. The work that follows builds a model based on the identification of cyberbullying in text data using natural language processing and machine learning utilising data from two distinct types of cyberbullying, hate speech tweets from Twittter and comments based on personal assaults from Wikipedia forums. To determine the most effective strategy, three feature extraction techniques and four classifiers are examined. The algorithm offers accuracy levels over 90% for data from Tweets and accuracy levels above 80% for data from Wikipedia.

1. INTRODUCTION

Technology has merged more completely into our lives than ever before. as the internet has developed. Social networking is popular right now. But just like with everything else, miscreants will emerge occasionally late and occasionally early, but they will exist nonetheless.Cyberbullying is a widespread problem nowadays.

Social networking websites are great resources for interperson communication. Although social networking has gained popularity over the years, most people still use it in unethical and immoral ways.

This often occurs amongst teenagers or even young adults. Bullying one other online is one of their harmful behaviours.Online, it is difficult to tell if someone is expressing anything only for amusement or whether he may have other motives.

Often, with just a joke, "or don't take it so seriously," they will laugh it off.The use of technology to target, harass, threaten, or humiliate another person is known as cyberbullying. This online conflict often turns into threats against some people in real life. Suicide has been used by certain individuals. At the beginning, these actions must be stopped. Any measures could be taken to prevent this, for instance, if a person's tweet or post is deemed offensive, his or her account might be closed or suspended for a specific time.

A Description Of The Project: Research on Cyberbullying Incidents reveals that in a 2018 poll by Child Right and You, an NGO in India, 11.4% of 720 young people were victims of cyberbullying, and over half of them did not even address it to their instructors, parents, or guardians.

Cyberbullying affected 22.8% of internet users aged 13 to 18 who spent around 3 hours online each day, compared to 28% of those who spent more than 4 hours online each day.

Numerous other reports have indicated that cyberbullying is having a negative impact on society as a whole, and that young people between the ages of 13 and 20 are particularly affected in terms of their physical and mental well-being as well as their capacity for making decisions at work. Researchers contend that each nation should take this issue seriously and work to find a solution. In 2016, the Blue Whale Challenge tragedy caused several kid suicides in Russia and other nations. It was a connection between an administrator and a player in a game that expanded over many social networks. Participants are assigned specific duties over a period of fifty days. At first, they seem simple, like getting up at 4:30 in the morning or viewing a scary movie. However, they eventually turned to self-harm, which led to suicides. Later it was discovered that the administrators were kids between the ages of 12 and 14.

2. LITERATURE SURVEY

2.1 Existing System

- Hsien[1] used datasets from four websites and an approach that includes opinion mining, social network analysis, and keyword matching to get an accuracy of 0.79 and recall of 0.71.According to a notion put up by PatxiGal'an-Garc'a et al. [2], a troll (someone who participates in cyberbullying) on a social networking site may always maintain a real profile to observe how other users see the fake profile. To discover these profiles, they recommended utilising machine learning. A few profiles that are relatively similar to them were examined using the identification procedure.
- The procedure included selecting profiles to look at, extracting information from tweets, selecting profile features to use, and applying ML to determine tweet authors. 1900 tweets from 19 different accounts were included in the analysis. It correctly identified the author 68% of the time. It was then used in a Case Study at a Spanish school where it was required to locate the real owner of a profile among a group of students who were suspected of engaging in cyberbullying. The following strategy still has several shortcomings.
- For example, experts who can change writing habits and styles such that no patterns are discovered or situations where trolling accounts lack legitimate accounts to deceive such algorithms. To adjust to evolving writing styles, more potent algorithms will be needed.
- A collaborative detection strategy was proposed by Mangaonkar et al. [3] in which data and results are combined to provide results from several connected detection nodes using either the same or different algorithms. P. Zhou et al. [4] suggested a B-LSTM approach based on concentration. Banerjee et al. [5] used KNN and new embeddings to attain a precision of 93%.

2.2Proposed System

This project's solution to the challenge of detecting cyberbullying involves categorising content as either having or not including the two main types of cyberbullying: hate speech on Twitter and personal assaults on Wikipedia.

• Tokenization: Tokenization is the process of dividing unprocessed text into meaningful words or tokens. For instance, the phrase "we will do it" may be tokenized as "we," "will," "do," and "it." Sentence tokenization and word tokenization are two different types of tokenization. [6] [7]Although there are many more types of tokenization, we use the Regex Tokenizer for this project. A regular expression is used in the regex tokenizer to determine the tokens to be used. The following regular expression is used to select tokens: Eg All of the alphanumeric tokens are retrieved for the regular expression "w+."

• Stemming: The transformation of a word into a root word or stem is known as stemming. For instance, the stem for the phrases "eating," "eats," and "eaten" is "eat." Since the root word "eat" has three branches, all three of them should be understood to mean the same thing. Porter, Lancaster, Snowball, and Regexp stemmers are the four varieties of stemmers that NLTK provides. The project that follows use PorterStemmer.

• Stop word removal: Stop words are words that add no sense to a phrase, such as the terms what, is, at, and an in English. You may delete these words since they are unnecessary. A list of English stop words is included in NLTK and may be used to filter out all tweets. When we train deep learning and machine learning models, stop words are frequently removed from the text data because the information they provide is irrelevant to the model and aids in enhancing performance.

2.2 Twitter Database

Two datasets that include hate speech were merged to create the Twitter Dataset:

Love Speech There are 17,000 tweets in the Waseem, Zeerak, and Hovy, Dirk Twitter Dataset[11] that have been flagged as racist or sexist. The annotations are used to mine the tweets.Due to account deletion or account deactivation, 5900 tweets were lost.[8] [9]

Hate Speech Language Dataset by Thomas Davidson,[10] [11] Dana Warmsley, Michael Macy, and Ingmar Weber[12].It included 25,000 tweets gathered via crowdsourcing.

This results in a total of 35787 tweets for the task distribution shown in Fig. 3.In the dataset that follows, training data make up 70% (25,050) of the dataset, while testing data make up 30% (10,737).

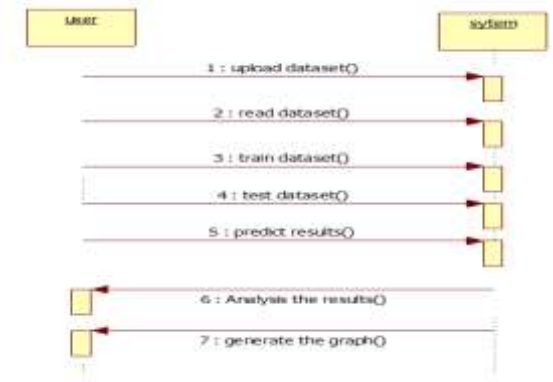


Fig 1: System Flow

3. SYSTEM ANALYSIS&DESIGN

3.1 Feasibility Study

Encyclopaedia data

The Wulczyn, Thain, and Dixon[13] Wikipedia dataset comprises 1M comments with the category "personal attacks."40000 comments from the dataset are utilised for the study, and 13000 of those comments have been classified as cyberbullying since they include personal attacks.[13][14] These remarks were taken from discussions among Wikipedia editors on articles that had been annotated by 10 different people using Crowd Flower. The similar split was used for this dataset (70 percent, or 28000, went to training data, and 30 percent, or 12000, went to testing data)[15].

3.2 Data Preprocessing

All text data is first changed to lowercase.Following that, some words like "what's" or "can't" are changed to "what is" or "can not."Additionally, the string library is used to remove all punctuation.The Natural Language Toolkit is then used to perform the following NLP techniques:

Tokenization: Tokenization is the process of dividing unprocessed text into meaningful words or tokens. For instance, the line "we will do it" may be tokenized into the words "we," "will," "do," and "it." Tokenization may be used to either words or phrases. The latter is known as sentence tokenization. Although there are many more types of tokenization, we use the Regex Tokenizer for this project. A regular expression is used in the regex tokenizer to determine the tokens to be used. The following regular expression is used to select tokens. Eg All of the alphanumeric tokens are retrieved for the regular expression "w+."

The process of stemming is turning a word into its basic word or stem. For instance, the stem for the phrases "eating," "eats," and "eaten" is "eat." Since the root word "eat" has three branches, all three of them should be understood to mean the same thing. Porter, Lancaster, Snowball, and Regexp stemmers are the four varieties of stemmers that NLTK provides. The project that follows makes use of PorterStemmer.

Stop words are words that contribute no sense to a statement. Some examples of stop words in the English language include: what, is, at, a, etc.

You may delete these words since they are unnecessary.

A list of English stop words is included in NLTK and may be used to filter out all tweets. When we train deep learning and machine learning models, stop words are frequently removed from the text data because the information they provide is irrelevant to the model and aids in enhancing performance.

3.3 System Architecture:Feature Extraction

Natural Language Processing needs feature extraction.Texta data must be transformed into numerical data since they cannot be categorised by classifiers. Each piece of information (in this example, a tweet or a remark) may be expressed as a vector, and such vectors can be utilised for categorization. The study that follows investigates three techniques for feature extraction:

a Bag of Words design

The BoW, or bag of words model, is a straightforward technique for extracting characteristics from texts that makes use of word occurrences. The Bag of Words model includes two crucial components:A list of words (or tokens) drawn from all documentsA method for evaluating each document's qualities using all of these phrases.It is called "bag" because the model simply considers the word itself, not how many times it appears in the text. This approach is predicated on the idea that comparable papers will include similar terms.The Bag of Words model follows the following process:Each document is used to create a vocabulary. All words (tokens) in all texts may make up the vocabulary, or just some of the tokens with the highest frequency, such as the top 10 characteristics with the greatest number of occurrences in the corpus. Additionally, vocabulary features can be extracted in a variety of ways depending on the number of words used for each feature. for the phrase "This was the best ever," for instance.

- One-word phrases like "this," "was," "the," "best," and "ever" are examples of a unigram model utilised in the corpus.
- Bigram model employs two words at a time for a feature, for example, "this was," "was the," "the best," and "best ever."The generalised model known as the "N-gram model" allows for the possibility of more than one value of N, such as the extraction of all unigram and bigram characteristics.

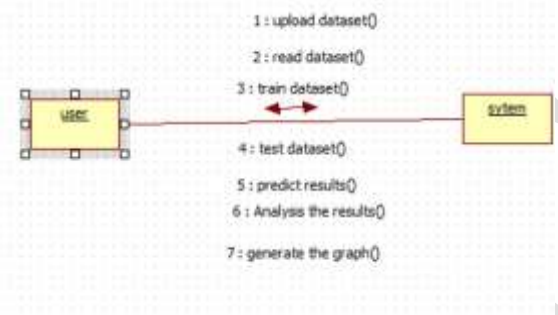


Fig 2 : System Architecture

3.4 Data Flow Diagram : Whenever a new system is developed, user training is required to educate them about the working of the system so that it can be put to efficient use by those for whom the system has been primarily designed. For this purpose the normal working of the project was demonstrated to the prospective users. Its working is easily understandable and since the expected users are people who have good knowledge of computers, the use of this system is very easy.

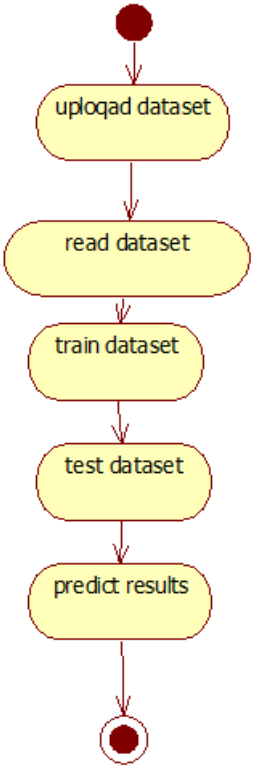


Fig 3: Data Flow Diagram

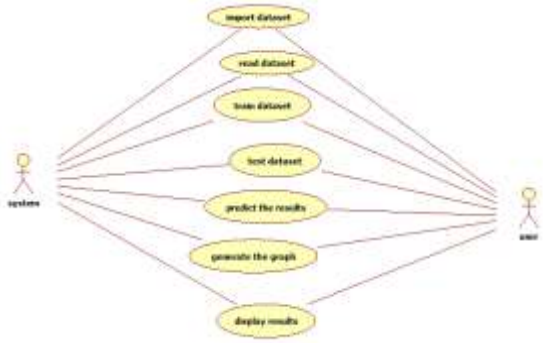


Fig 4 Use Case UML Diagrams

4. IMPLIMENTATION AND RESUTS

Review of Literature and Open Challenges for Machine Learning Algorithms Predicting Cyberbullying on Social Media in the Big Data EraIn this study, the author predicts cyberbullying postings from social media using a variety of machine learning techniques, including SVM, Random Forest, Naive Bayes, KNearest Neighbours, and Decision Tree. 'Extreme Machine Learning' algorithm, a cutting-edge algorithm in the field of machine learning, is used as a second algorithm.

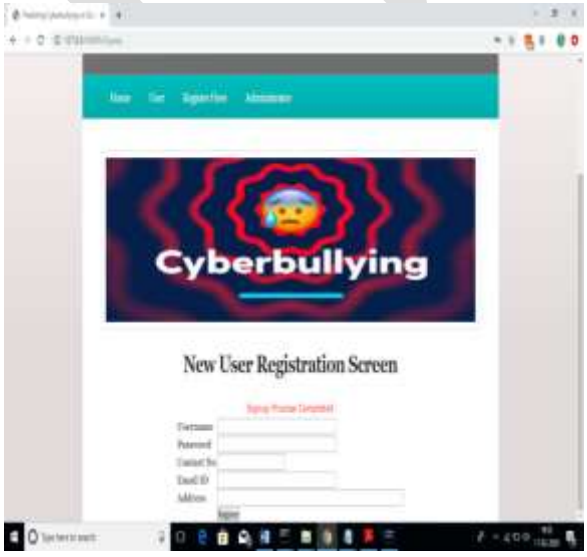
We will create a train model with normal and bullying messages using all available algorithms, and this model will be used to new user posts to determine if they are normal or include bullying material.There are the following modules in this project.

Users may create accounts using the user module. They can log into the application with their account information and then send and view posts.

Administrator Module: The administrator may examine all users who have registered and can then approve or disapprove new users. New bullying messages should be added to the machine learning training dataset by the administrator. In order to accomplish bullying message identification from the user side, the administrator must run all or at least one SVM algorithm. All posts sent by all users may be seen or tracked by the admin.

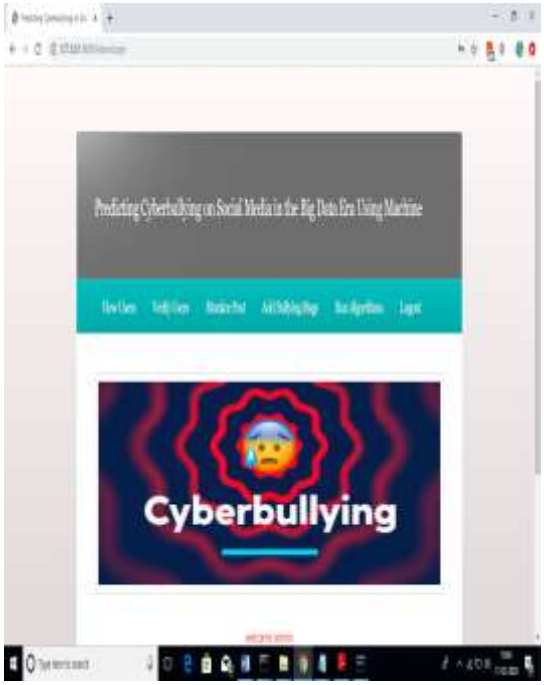
Note that if just a little amount of relevant data is included in the train model, machine learning algorithms will predict whether a message will be bullied or not. As a result, you can accurately anticipate every message that the administrator enters. The 'Cyber/dataset.txt' file contains all example anti-bullying and pro-bullying texts.Create a database first by copying and pasting the contents of the "DB.txt" file into the MYSQL console.Install DJANGO, deploy the "Cyber" folder, start the server, and open the browser by typing the URL "http://127.0.0.1:8000/index.html" to execute this project. Run the aforementioned URL to see the screen below.

In above screen click on ‘Register Here’ link and add new user to create account



In above screen sign up process completed. Now click on 'Administrator' link to login as admin and give permission to new user

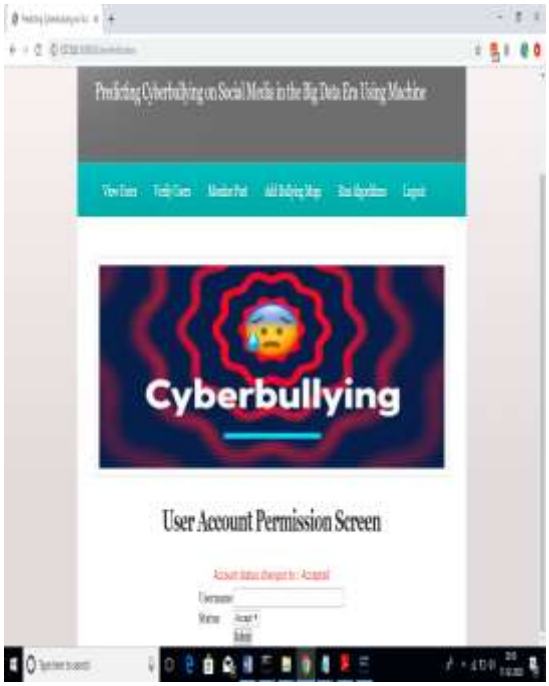
In above screen login as 'admin' by giving username as 'admin' and password as 'admin'. After login will get below screen



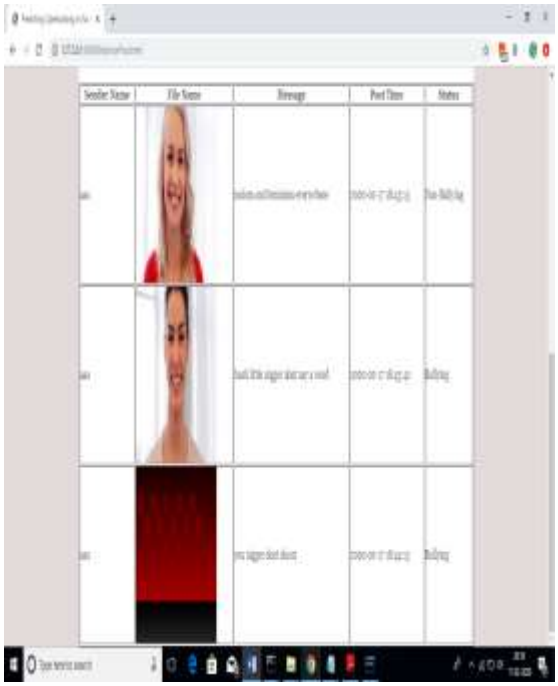
Now admin can click on 'View Users' link to view all users list

In above screen we can see 'rajesh' account is in pending state and to give permission to rajesh. Now admin will click on 'Verify Users' link to get below screen and to give permission

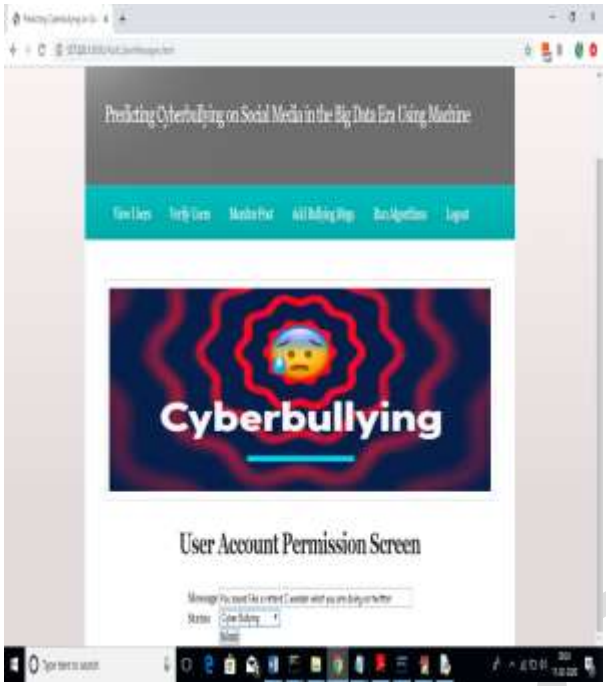
In above screen admin will enter username and then select 'Accept' or 'Reject' option to give permission.



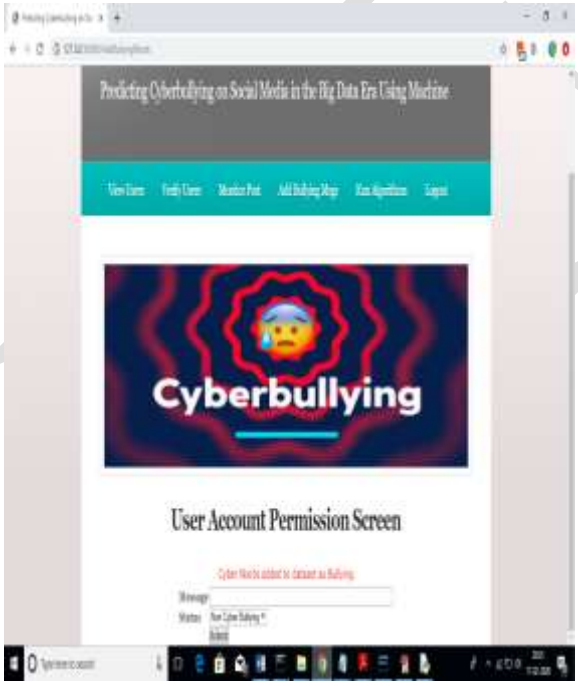
In above screen account state changed to 'Accepted'. Now admin can click on 'Monitor Post' to view all post from past users



In above screen application will automatically detect whether message is non-bullying or bullying from machine learning algorithms. Now admin can click on 'Add Bullying Msgs' link to add words

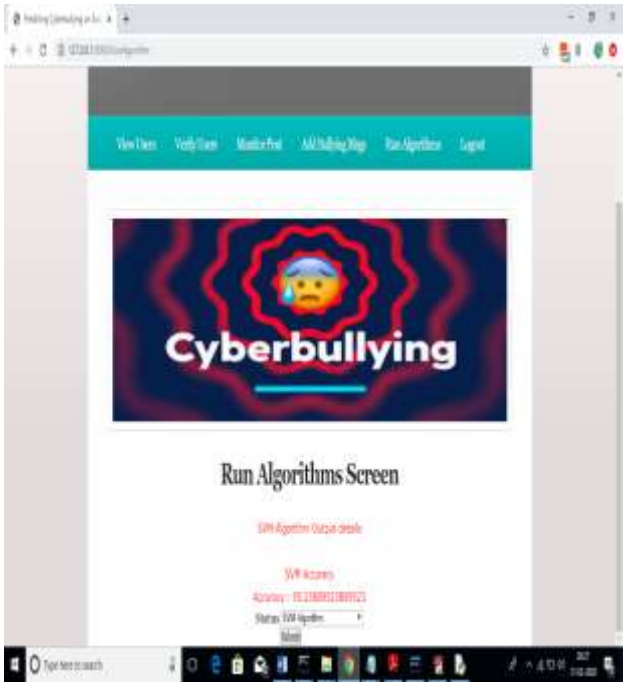


In above screen admin adding one sentence as 'Cyber Bullying' and similarly he can add all possible bullying and non-bullying messages. After adding messages will get below screen

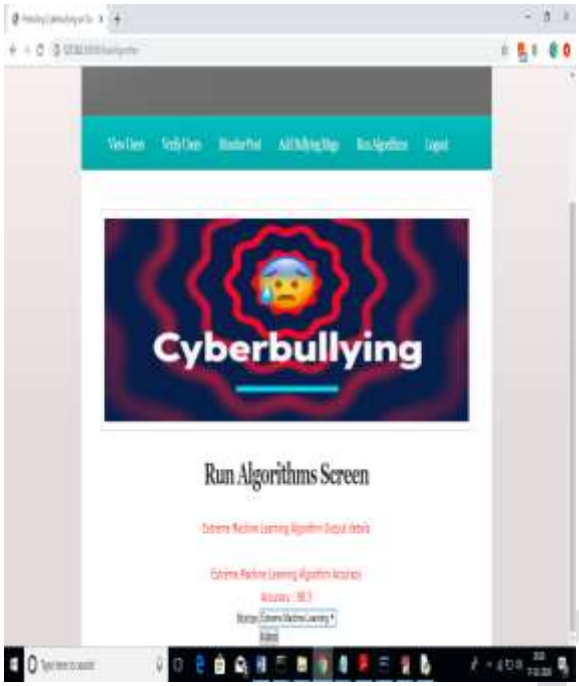


Now admin can click on 'Run Algorithms' link to generate train model using entire dataset to predict user posts as normal or bullying

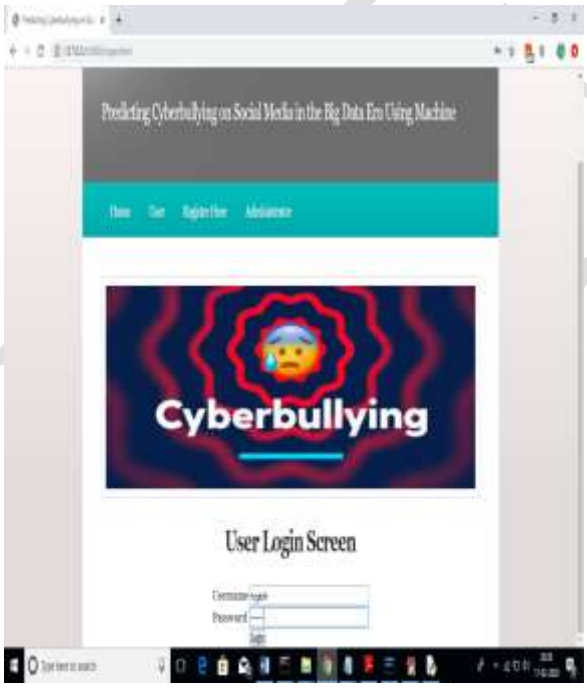
In above screen admin has to select each algorithm and click on 'Submit' button to train model and we will get accuracy also for each algorithm. Admin has to repeat this step whenever first time he starts the server or upon adding new bulling messages.



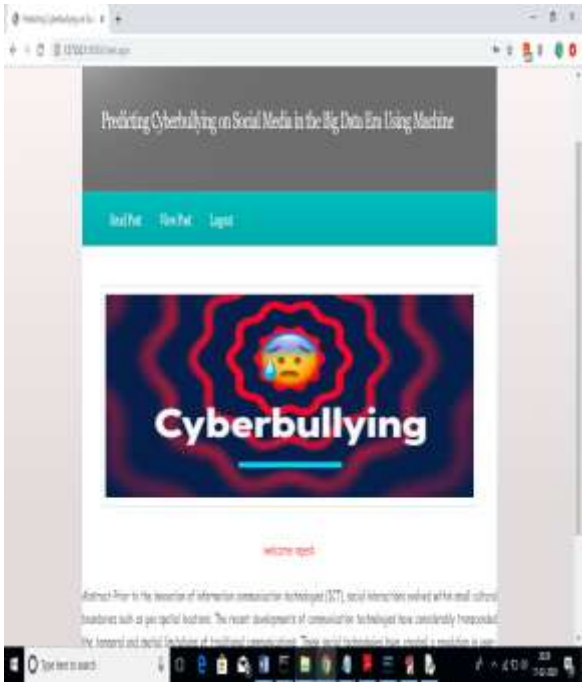
In above screen I ran SVM and got accuracy as 95. Similarly u need to select all algorithms one by one and run it.



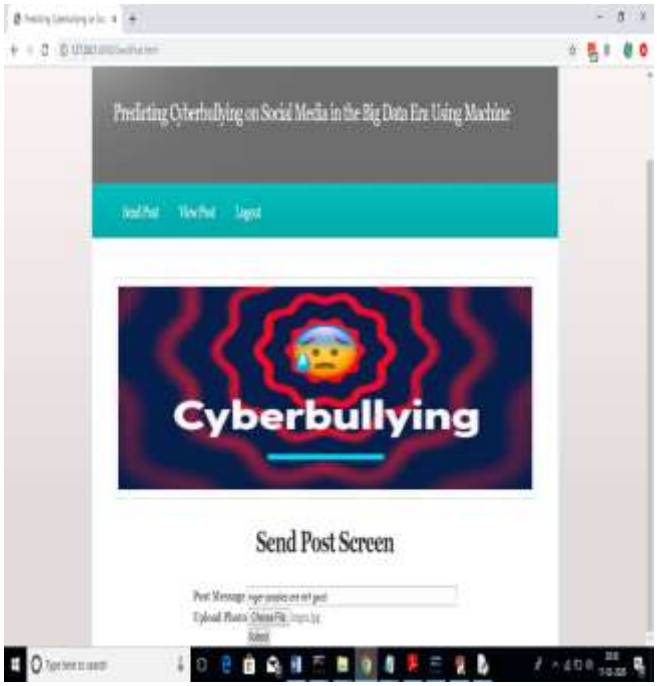
In above screen advance 'Extreme Machine Learning' algorithm gave 98% accuracy. Now admin logout and login as user to send posts.



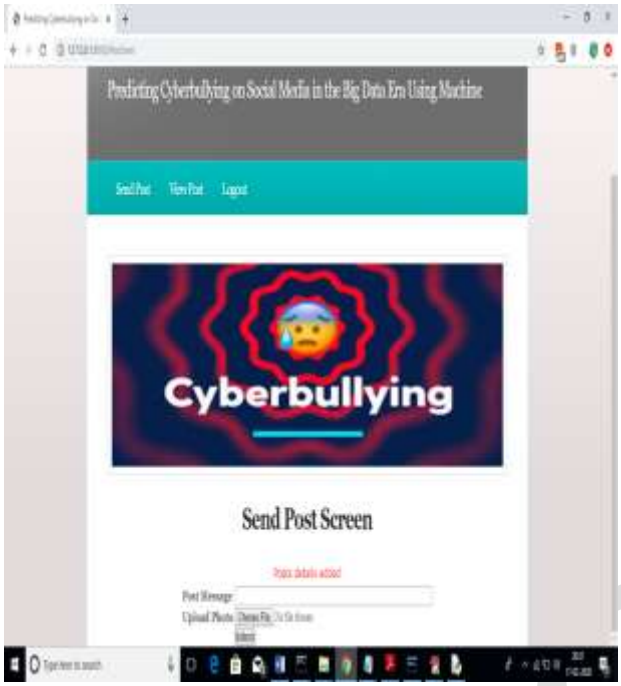
In above screen rajesh user is login and after login will get below screen



In above screen click on 'Send Post' link to get below screen



In above screen as post I added some messages and uploaded a photo also. After posting message will get below screen



In above screen we are seeing posts from all users and rajesh post predicted as 'Non-Bullying'. Here based on words given in dataset will get prediction as bullying on non-bullying.

5. CONCLUSION

There is a need to prevent the growth of cyberbullying since it is hazardous and causes unfortunate events like suicides and sadness. Consequently, it is crucial to detect

cyberbullying on social media platforms. More data and better-classified user information are now available for numerous additional types of cyberattacks. On social media platforms, cyberbullying detection can be used to block users who attempt to engage in such behaviour. In this research, we suggested a detection architecture for cyberbullying to address the issue.We spoke about the data architecture for hate speech on Twitter and personal assaults on Wikipedia.Given that tweets containing hate speech often included profanity, which made it simple to identify, natural language processing approaches for this kind of speech were successful with accuracy rates of over 90% utilising fundamental machine learning algorithms.Because of this, it performs better with BoW and Tf-Idf models than Word2Vec models.

Although the three feature selection methods performed similarly, it was challenging to identify personal attacks using the same model because the comments lacked a lot of learnable sentiment.When integrated with Multi Layered Perceptrons, Word2Vec models that exploit the context of features gave comparable results in both datasets with significantly less features.

6. FUTURE SCOPE

In the near future, we expect to extend this research in several ways. First, we would like to explore heterogeneous combinations of different machine learning models on the dual model. For example, Extra Tree for the positive model, while Random Forest for the negative model. Second, we plan to further extend the features regarding comments, as these concentrate most of the information for the early detection. Third, we would like to investigate an evaluation based on time, instead of number of posts, since it may be relevant for the early detection of cyberbullying. Finally, we intend to experiment with other datasets from some other social media platforms to validate our approach and generalize the results.

7. REFERENCES

[1] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, “Towards the detection of cyberbullying based on social network mining techniques,” in Proceedings of 4th International Conference on Behavioral, Economic, and SocioCultural Computing, BESC 2017, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.

[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, “Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying,” 2014, doi: 10.1007/978-3-319-01854-6_43.

[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, “Collaborative detection of cyberbullying behavior in Twitter data,” 2015, doi: 10.1109/EIT.2015.7293405.

[4] R. Zhao, A. Zhou, and K. Mao, “Automatic detection of cyberbullying on social networks based on bullying features,” 2016, doi: 10.1145/2833312.2849567.

[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, “Detection of Cyberbullying Using Deep Neural Network,” 2019, doi: 10.1109/ICACCS.2019.8728378.

[6] K. Reynolds, A. Kontostathis, and L. Edwards, “Using machine learning to detect cyberbullying,” 2011, doi: 10.1109/ICMLA.2011.152.

[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020, doi: 10.1109/ICESC48915.2020.9155700.
[8] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," arXiv. 2018.
[9] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," arXiv. 2018.
[10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards the identification of cyberbullying in social networking sites," 2016, doi: 10.1109/ASONAM.2016.7752420.
[11] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," 2016, doi: 10.18653/v1/n16-2013.
[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," 2017.
[13] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," 2017, doi: 10.1145/3038912.3052591. [14] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," Artif. Intell. Rev., vol. 53, no. 6, 2020, doi: 10.1007/s10462-019-09794-5.
[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.
[15] Shantala S Yalawar K Vijaya Babu, Mrutyunjaya S Yalawar, A New Approach for Secure Login Method and Forestall Cyber bulling in Social Media, 2019/5, International Journal of Recent Technology and Engineering (IJRTE),Volume 8, Issue Ic2,1243-1246. https://scholar.google.co.kr/citations?view_op=view_citation&hl=en&user=Tyqm5YUAAAAJ&citation_for_view=Tyqm5YUAAAAJ:UebtZRa9Y70C