WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Cyber Bulling Detection in Chat Application

***Ms.N.Dharani*[1], *Mr.P.Anbumani.,MCA.,M.Phil.,NET.*[2], *Mr.Sivakumar,M.E.,*[3]**

[1]Masters of Computer Applications, Krishnasamy College of Engineering and Technology, Cuddalore, TamilNadu, India
[2]Associate professor,Masters of Computer Applications ,Krishnasamy College of Engineering and Technology, Cuddalore, TamilNadu, India.
[3]Human Resource,Photon Technologies,Cuddalore,TamilNadu, India

ABSTRACT

Social Network Services (SNS) is an online platform where teenagers/young people who are addicted to social media are more likely to engage in cyberbullying. An activity of threatening, insulting, and bullying a person through messages comes under cyberbulling. As messaging apps are increasing, cyberbullying is rising day by day. Cyberbullying is a misuse of advanced technology to persecute a person. To preclude cyber victims from the activities is challenging. However, many social media bullying detection techniques have been implemented but not automatic detection in a live chat application. A project aims to detect horrifying words/ hazardous content in live Chat applications/Boards. Two classifiers i.e., the Naïve Bayes classifier and logistic regression classifier are used for training and testing the dataset to predict abusive words in live conversation.

Keywords – Cyberbullying; Machine Learning; Cyberbullying Detection; Systematic-Review; Prevention.

I. Introduction

Due to the development of the internet, technology and Social Network services, such as Facebook, Twitter, Instagram, and WhatsApp are acquiring in popularity as the main source of spreading messages to other people. Generally, Communication happens through messages and is very useful in various sectors, for example, business, education, and socialization. However, it also provides an opportunity to create harmful activities. There are numerous shreds of evidence showing that messaging can introduce a very concerning problem, namely cyberbullying.

Cyberbullying involves the offensive information in the messages which are sent using SNS to intentionally hurt people emotionally, mentally, or physically. It can cause manifest psychological problems such as loneliness, low self-esteem, social anxiety, depression, and a variety of other emotional problems, and even leads to suicide. It is tragic consequences have been continuously reported typically about 36% of children in India. Since the number of cyberbullying incidents has recently been raising, an intensive study of how to effectively detect and prevent it from happening in a real-time environment is needed. To preclude victims from the incidents, blocking the message is not an effective way to tackle cyberbullying. Instead, text messages should be monitored, processed, and analyzed as quickly as possible to support real-time decisions.

As the problems mentioned, some studies are assigned to explore various techniques to detect cyberbullying efficiently. Manual detection is considered the most accurate detection, but it is hardly employed because it takes too much time and lots of resources. An automatic cyberbullying detection system is therefore emphasized.

Even though cyberbullying detection system has extensively been explored, cyberbullying remains a growing concern and the existing approaches are still inadequate, especially when dealing with a huge volume of data. Various kinds of SNS can represent different forms or patterns of data. The problem can be defended by detecting and preventing it by using a supervised machine learning approach which can be done from different perspectives. The main purpose of our paper is to build a classification model to predict the text messages for preventing cyberbullying in a live environment. Furthermore, the Detection process is automatic, abusive words are detected fast, and the result warning message is displayed which automatically blocks the user who used the abusive word.

II. Literature Survey

In recent years, several studies on online bullying analysis, detection and prevention using text mining by classifying conversations have been published. Detecting social media bullying is done by John Hani et al.[1]. In their paper, to detect and prevent social media bullying they have been used Neural Networks and SVM to build classification model. For the proposed model they collected the dataset from the Kaggle. The proposed model is divided into 3 major steps:

- Preprocessing Steps:
    - Tokenization
    - Lowering Text
    - Stop words
    - Word correction

- Feature Extraction: For **feature extraction** sentiment analysis and TF-IDF algorithms are used.
- Classification: For **classification** SVM (Su

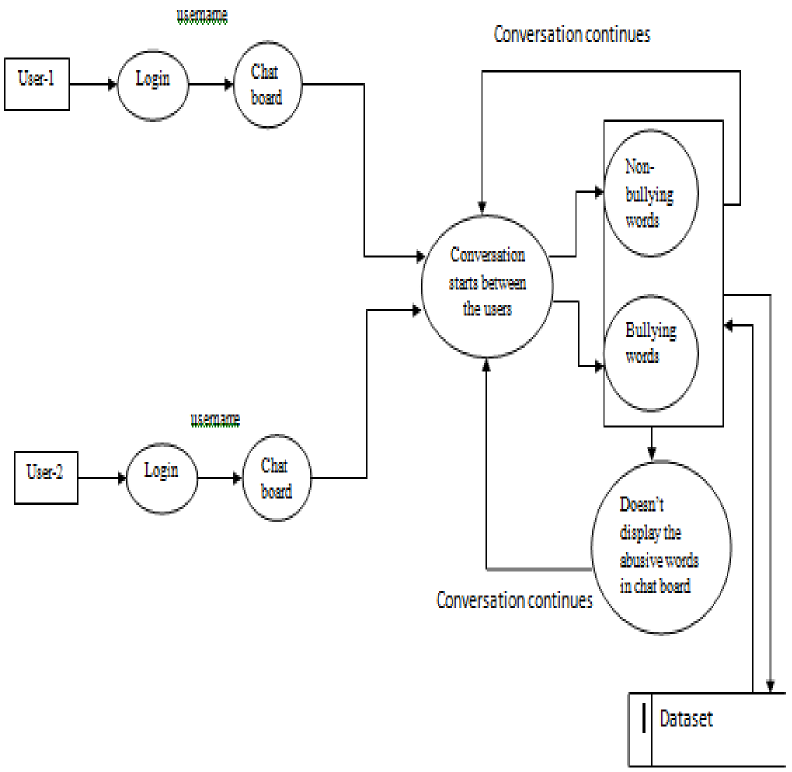
Kelly Reynolds et al. [2] has proposed a ma ... om
Formspring.me website where users ask and answe ... and
eleven classes will be classified in this research to m ... are
mostly used .The class label “no” and “yes” for a twe ... one
for counting information and one for normalizing th ... and
k-nearest neighbor (k = 1 and k = 3).

Amanpreet Singh et al. [3] has reviewed many p ... ine
learning models, etc. This paper includes study res ... ets,
conclusions/findings, content-based features, demer ... ses,
they've explored Scopus and the IEEE Xplore virtu ... on
concluding arguments, abstracts, and titles, 18 pape ... ey,
they've
reviewed 27 papers from 33 papers after filtration. In ... most
of them have used the Support Vector Machine (SV

Karthik [4] applied multi-classifiers suc ... act
features and employ the weighted term frequency- ... om
Kongregate, Slashdot, MySpace and Homa [6] used

Support Vector Machine while using Din ... nts,
manually label them, and implement various bina ... in
classification, Cynthia Van Hee, Els Lefever, Ben Ve ... ch,
when the preprocessing step occurs, they use the Le ... on
those facts, this study will be conducted to classify ... lds'
previous research (2012).

This project will be built with Python and socl ... the
model. After pre-processing, the dataset is then trair ... ms.
After that,, we will create a web-based application ... to
determine whether the messages are bullying or not.



III. Proposed system

An automatic cyberbullying detection system is to detect, identify, and classify cyberbullying activities from the large volume of streaming texts from Live chatting. For each message, cyberbullying is detected using the model and then alert messages are posted on chat boards. Texts are fed into the cluster and discriminant analysis stage which can identify abusive texts. The abusive texts are then clustered by using Naïve Bayes is used as classification algorithms to build a classifier from our training datasets and build a predictive model. The first method aims to clean and pre-process our datasets by removing non-printable and special characters, reducing the duplicate words, and clustering the datasets. The second one concerns the classification model to predict the text messages for preventing cyberbullying.

1) Data Collection:

The data used to create a data set is a textual conversation taken from the online site -Kaggle (www.kaggle.com) which provides 2,000 conversations. The question, Answer, and Severity is the fields used as a label in this research. Each conversation is a combination of Question and Answer fields. The combined results of Q&A from excel files are made into files with .txt and grouped in folders 0 through 10 according to the severity level used as labels. After data collection, data is imported into Rapid Miner to continue the process of Preprocessing, Extraction, Classification, and Evaluation.

2) Preprocessing:

Conversation Text on each set of data is later preprocessed to facilitate the processing of text conversations at the next stage:

i. Data Cleaning & Data Balancing:

The amount of data obtained from www.kaggle.com is 12,729 data, including 11,661 data given a non- cyberbullying label and 1068 data labeled cyberbullied. Data cleaning is done with MS-excel by removing conversations that have total characters under 15 letters, deleting meaningless words like "haha", "hehe", "uh", "hmm", "umm". For data balancing on the classification of 2 classes cyberbully, non-cyberbully), 4 classes (non-cyberbully, cyberbully level severity low, cyberbully level severity middle, cyberbully level severity high), and 11 classes (non-cyberbully, cyberbully level severity 1 – 10), then the data used amounted to 1.600 for balancing data.

ii. Tokenization:

Tokenization is the process of cutting or separating each word that compiles a document or conversation. In general, every word is identified or separated from other words by a space character, single quoting character ('), dot (.), semicolon (;), colon (:), so the tokenizing process uses nonletters mode to perform word separation.

iii. Transform case:

Transformation into the lower case to facilitate the next process with the purpose of not distinguishing between capital letters and lowercase letters.

iv. Stop Word Removal:

Delete unnecessary words in every text conversation under English vocabulary by using Stop Word Filter (English). 5) Filter Token: The token filter is selecting the word that the number of characters between 3-25, because below 3 characters word is a stop word and above 25 are character is rarely used words.

v. Stemming:

The words in the text conversation are transformed into basic words using the Porter Stemmer algorithm.

vi. Generate n-grams:

The process of generating n-grams is to form a set of words from a parable and graph, usually by moving one word forward, in this research an n-gram of 2 to 6, because the experiments have been done n-gram over 6 is stable (the result is the same as n-gram).

3) Extraction:

The preprocessed conversations will be transformed into a vector model where text conversations are represented with a vector of extracted features. Features resulting from the extraction are words or combinations of words to form a list of words and the calculation of the weight with a count vectorizer. In this stage, the classification willuse the Naïve Bayes method to generate the model.

4) Classification:

In machine learning , Classification[8] is a supervised learning approach in which a computer program learns from given data and makes new observations or classifications. It is a process of separating a given set of data into classes which can be performed on both structured and unstructured data and these classes are often referred to as target, label or categories. The main aim is to predict in which category the new data will falls into.

The most common problems are – speech recognition, face detection, handwriting recognition, document classification, etc. For example, when filtering emails “spam” or “not spam”, when looking at transaction data, either “unauthorized”, or “authorized”. There are several classification models which includes logistic regression, decision tree, random forest, and Naive Bayes.

1. Naive Bayes Model:

Naive Bayes is a supervised probabilistic machine learning algorithm that can be used for classification that work by applying Bayes theorem with naive independence assumptions between the different features. Naive Bayes models are used for recommendation systems, sentiment analysis, and spam filtering and very easy to implement.

The Naive Bayes classifier needs a small amount of training data to estimate the necessary parameters to get the results and they are extremely fast as compared to othe r classifiers. The only demerit is that they are known to be a bad estimator.

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B = events

$P(A|B)$ = probability of A given B is true

$P(B|A)$ = probability of B given A is true

$P(A), P(B)$ = the independent probabilities of A and B

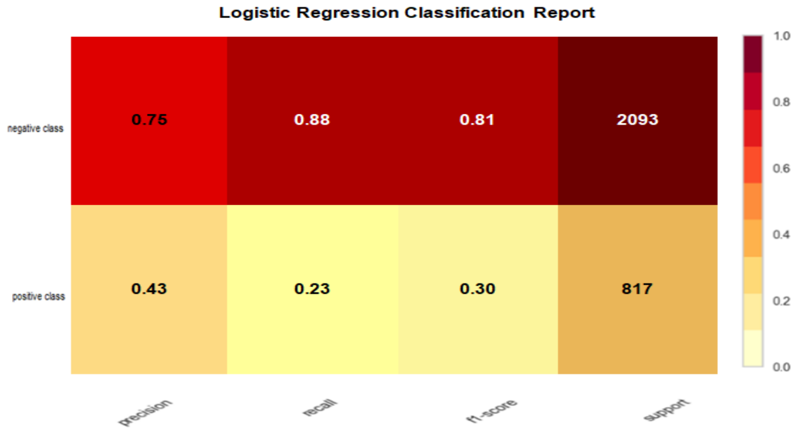
It assumes that the existence of a particular feature in a class is not related to the existence of any other feature. A very simple document representation is used here, usually a bag of words. In the case of severity, words are very important for the meaning of the text, and thus imperative in its classification, are considered and given weight according to meaning,. For instance, "faggof” would receive a higher weight than "bitch", due to the former being sexually discriminatory and abusive.

2. Logistic Regression:

It uses the concept of pred sion, but is used to classify samples; therefore it falls under the classificati 0 and 1, which cannot go beyond this limit, such as values above the thresho vhich forms a curve like the "S" form. This S-form curve is called the Sigm

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In logistic regression, we will do scaling for feature because for prediction we need the accurate result. Now we will train the dataset using the training set. We will import the *LogisticRegression* class of the *sklearn* library for detetction. After importing the class, we will create a classifier object to fit the model.



IV. System architecture

To detect live chat bullying automatically, supervised classification machine learning algorithms like Logistic Regression and Naive Bayes is incorporated. The reason behind this is both Logistic Regression and Naive Bayes calculate the probabilities for each class (i.e. probabilities of Bullying and Non-Bullying Messages). Both Naive Bayes and Logistic Regression algorithms are used for the classification of the two-cluster were evaluated on the same dataset.

Classification report also evaluated and the accuracy, recall, f-score, and precision are also calculated.

Precision = TP / (TP+FP)

Recall =TP/(TP+FN)

F-Score = 2*(Precision*Recall) / (Precision + Recall)

Where TP = True positive numbers

TN = True negative numbers

FN = False negative numbers

FP = False positive numbers

This project contains the following modules:

Model Training Module:

In this Module, the data set is collected and data is pre processed and then converted using a count vectorizer. The Testing training data set is divided and the algorithm is initialized. Features and labels are fitted into the algorithm. The model is saved to the system after being predicted with accuracy.

Server Module:

Server Module has socket programming where the port and IP address are connected to manage messaging by communicating with clients and loads trained model to check each message and detect if bullying words are used and then the message is sent to clientUI.

Client Module:

The client module is designed using the Tkinter framework which is connected to IP and port number. Chats and messages with other clients are viewed from the server to detect if there is any usage of unauthorized words.

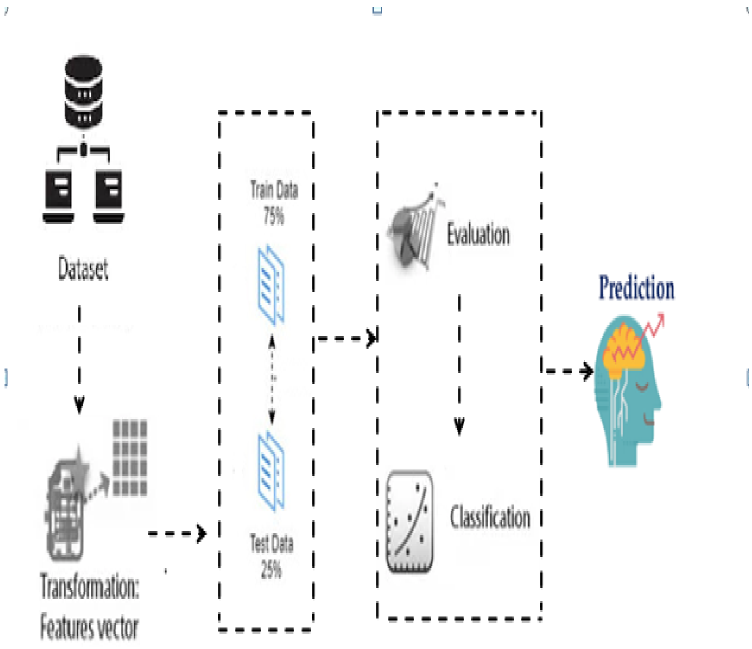
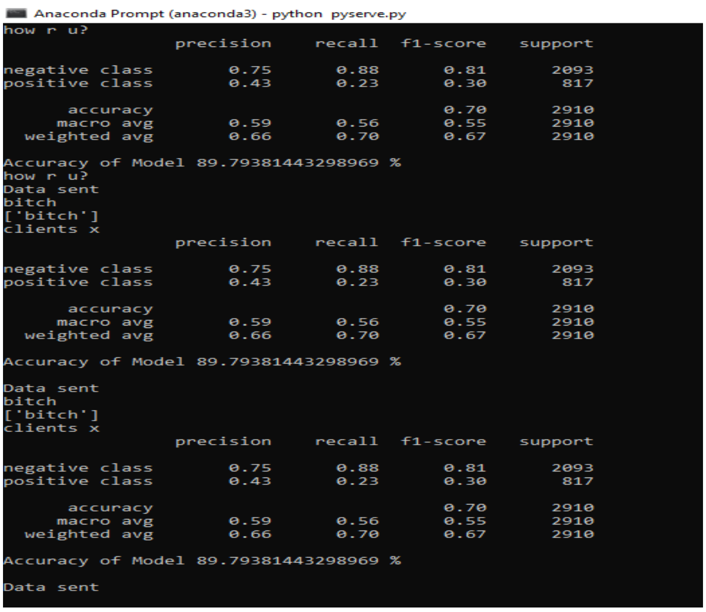


Fig: Architecture Diagram

V. result

The theme of the project is to maintain peace and contribute to society with the help of trending and emerging technology i.e. Machine Learning. We had perfectly utilized the features of the learning. This system is used to detect and prevent abusive conversations between the chat during live conversations. The model is developed using a Naïve Bayes classifier for evaluating the classifier and logistic regression classifier for detecting abusive words. The combined technology of ML with Python is being used to train and test the model with high accuracy. The model provides some features which can acknowledge the abusive words and replace them with asterisks which will automatically block the person thereby preventing cyberbullying.



```
Anaconda Prompt (anaconda3) - python  pyserve.py
how r u?
                precision    recall  f1-score   support

negative class       0.75      0.88      0.81      2093
positive class       0.43      0.23      0.30       817

      accuracy                           0.70      2910
     macro avg       0.59      0.56      0.55      2910
  weighted avg       0.66      0.70      0.67      2910

Accuracy of Model 89.79381443298969 %
how r u?
Data sent
bitch
['bitch']
clients x
                precision    recall  f1-score   support

negative class       0.75      0.88      0.81      2093
positive class       0.43      0.23      0.30       817

      accuracy                           0.70      2910
     macro avg       0.59      0.56      0.55      2910
  weighted avg       0.66      0.70      0.67      2910

Accuracy of Model 89.79381443298969 %

Data sent
bitch
['bitch']
clients x
                precision    recall  f1-score   support

negative class       0.75      0.88      0.81      2093
positive class       0.43      0.23      0.30       817

      accuracy                           0.70      2910
     macro avg       0.59      0.56      0.55      2910
  weighted avg       0.66      0.70      0.67      2910

Accuracy of Model 89.79381443298969 %

Data sent
```

fig: Classification report

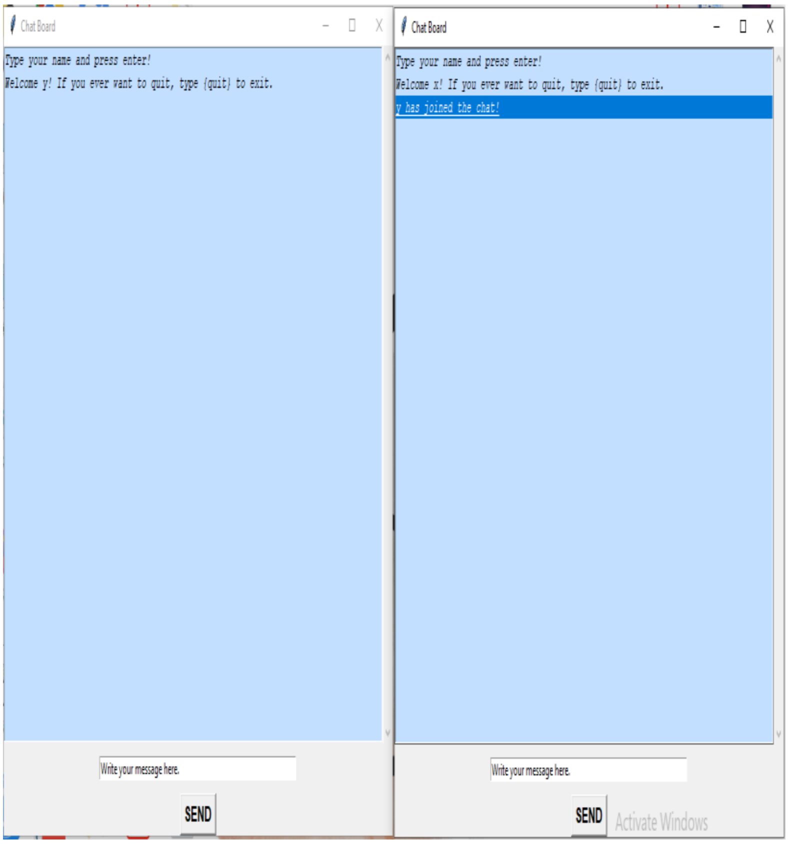


Fig: Chatting between 2 clients

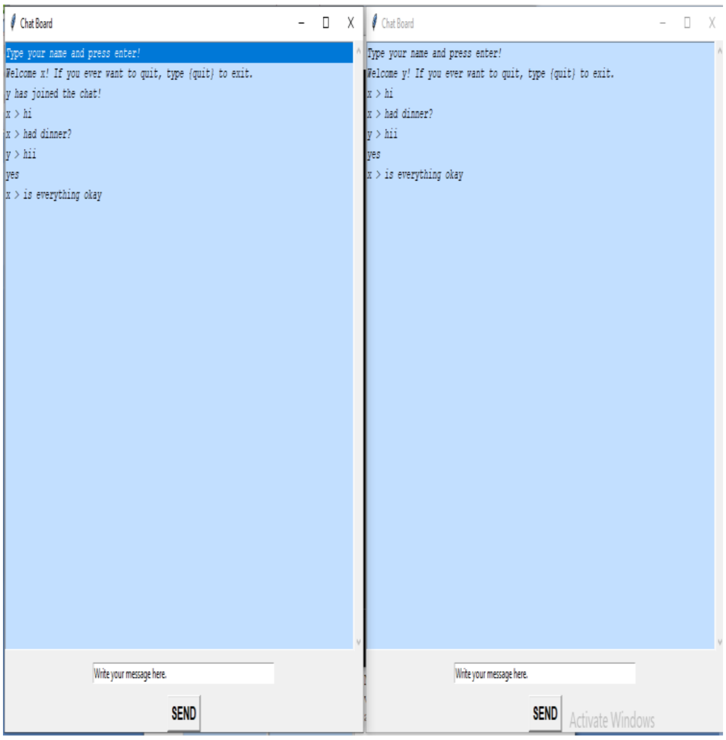


fig: Hides the Bullying words

VI. Conclusion

Automatic Cyberbullying detection is a crucial task in Live Chat applications. In this paper, an approach is proposed for detecting and preventing bullying using Supervised Machine Learning Algorithms. Our proposed classification model is evaluated on both Logistic Regression and Naïve Bayes for feature Extraction. As a result, the accuracy of detecting cyberbullying content of around 89.79 % which is better than Support Vector Machine. This model will help people to ***detect horrifying words and*** hazardous words or content ***in live Chat applications/Boards.***

VII.FUTURE ENHANCEMENT

An interesting direction for future work would be the detection of Audio cyberbullying message categories such as threats, curses, and expressions of racism and hate using sentiment analysis. When applied in a cascaded model it could find critical cases of cyberbullying with high precision which will be more helpful for monitoring purposes.

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely fortunate to have got this all along the completion. Whatever I have done is due to such guidance and assistance. We would not forget to thank them. I thank **Mr. P.Anubumani** and **Mr.SivaKumar** for guiding us and providing all the support in completing this project.

References

[1] John Hani Mounir, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad,Eslam Amer, Ammar Mohammed, " Social Media Cyberbullying Detection using Machine Learning", (IJACSA) International Journal of Advanced Computer Science and Applications Vol. 10, pages 703-707, 2019.

[2] Kelly Reynolds, April Kontostathis,Lynne Edwards, "Using Machine Learning to Detect Cyberbullying", 2011 10th International Conference on Machine Learning and Applications volume2,pages 241-244.IEEE,2011.

[3] Amanpreet Singh, Maninder Kaur, "Content -based Cybercrime Detect ion:A Concise Review", Internat ional Journal of Innovat ive Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8, pages 1193-1207, 2019.

[4] D. Karthik, R. Roi, and L. Henry, "Modeling the detection of textual cyberbullying," International Conference on Weblog and Social Media - Social Mobile Web Workshop, 2011.\

[5] N. Vinita, L. Xue, and P. Chaoyi, "An Effective Approach for Cyberbullying Detection," Communications in Information Science and Management Engineering, 2013, vol. 3, no. 5, pp.238-247.

[6] H. Homa, A. M. Sabrina, I. R. Rahat, H. Richard, L. Qin, and M. Shivakant, "Detection of Cyberbullying Incidents on the Instagram Social Network," 2015.

[7] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proc. IEEE International Fifth International AAAI Conference on Weblogs and Social Media (SWM'11), Barcelona, Spain, 2011.

[8]https://www.edureka.co/blog/classification-in-machine-learning/#:~:text=In%20machine%20learning%2C%20classification%20is,recognition%2C%20document%20classification%2C%20etc.

[9] https://www.javatpoint.com/logistic-regression-in-machine-learning#:~:text=Logistic%20regression%20is%20one%20of,of%20a%20categorical%20dependent%20variable.