# Detecting the Presence of Cyberbullying using Machine Learning

Arathi Unni
Department of Computer Science and Engineering
College of Engineering Kidangoor
Kottayam, KERALA

Ranimol K R
Department of Computer Science and Engineering
College of Engineering Kidangoor
Kottayam, KERALA

Linda Sebastian
Assistant Professor
Department of Computer Science and Engineering
College of Engineering, Kidangoor
Kottayam, KERALA

Rajalakshmi S
Department of Computer Science and Engineering
College of Engineering Kidangoor
Kottayam, KERALA

Sissy Siby
Department of Computer Science and Engineering
College of Engineering Kidangoor
Kottayam, KERALA

***Abstract*—The increasing use of online social media and their demand has turned up the rise of cyberbullying among people. Nowadays cyberbullying has become very frequent. The majority of the people are using social media to troll and smear others, and the others are being defamed and agitated by unknown users or friends. So it is necessary to detect these type of comments and prevent it. Our work proposes an ensemble learning approach to detect cyberbullying comments. Different supervised ensemble learning techniques are used to classify comments. Here voting classifier trains on an ensemble of Support Vector Machine, Logistic Regression, and Perceptron models and predicts the output based on the highest majority of the vote. This model detects cyberbullying comments with 94% accuracy.**

***Keywords—Cyberbullying; ensemble learning; Perceptron; Logistic Regression; support vector machine; voting classifier***

## I. INTRODUCTION

The real world stopped and the reel world started what could be more appropriate to describe the life of today's generation. The reel world works as a web-connected people from all over the globe. The medium which should have been used to connect and communicate with people has started to become a place where people, especially teens and young adults bully one another. Though the positive side where social media helps people to connect, it also exposes them to a threatening situation like aggressive cyberbullying which can lead to signals of depression and suicidal thoughts. So in the era of widespread usage of social networking, cyber-attacks are also on the rise. Cyberbullying is at the forefront of this, which is the act of insulting or defaming a person personally. Moreover, it includes harassment and flaming communications. Through text messaging, comments, etc. by cyber communication, cyberbullying can occur. Different models are generated to resolve this issue by detecting cyberbullying and a lot of detection has been done using many techniques. But still, cyberbullying is seen as a major issue on many social media.

Our work focuses on an ensemble learning approach. Machine learning is the technique that learns from several data and builds up a model that automatically classifies the specific action. It helps to detect language patterns and generate cyberbullying detection models. In this, we use a supervised machine learning approach to the data that are collected from online conversations. To train the model, we took datasets from sites such as Kaggle, YouTube, and Twitter. 80% of the data is held as the training data and the remaining 20% is held as the test data. We changed the dataset to a format that the model can easily work with it. Feature extraction and vectorization are done using TFIDF and n-gram to transform the text into feature vectors that can be inputted to three classifiers namely Perceptron, Support Vector Machine, and Logistic Regression. We also generated a voting classifier model that trains an ensemble of LR, SVM, and Perceptron models and predicts the output based on the highest majority of the vote. Finally, we evaluated the efficiency and accuracy of all the models with the same dataset.

The content of the paper is formulated as session 2 sights the literature review. Session 3 describes the methodology. Session 4 shows the analysis and results of the proposed approach. Session 5 describes the future scope and session 6 concludes the paper.

## II. LITERATURE REVIEW

Different methods are there that identify cyberbullying comments with high accuracy. Bandeh Ali Talpur [1] proposed a supervised machine learning technique for cyberbullying detection and multiclass categorization. They

applied sentiment, Embedding and Lexicon features with PMI-semantic orientation and applied the selected features to Decision Tree, SVM, KNN, Naive Bayes, and Random Forest algorithms. The developed model classified the Twitter contents as cyberbullying or non-cyber bullying along with the severity level as low, medium, high, or none. They found that Random Forest is the best classifier having an AUC of 0.971 and f-measure 0.929.

Another method by Risul and Nasrin [2] was to identify and filter cyberbullying comments in three different categories such as hate speech, offensive speech, and neither. Feature extraction methods such as bag-of-word, N-gram, and TF-IDF are used to create feature vectors and input data. They used logistic regression, random forest, and SVM models. During the training phase, they found that logistic regression and random forest have the same performance in classifying comments. After the training phase, the testing data is applied to the three models. They found that logistic regression classifies comments at 86% accurately, the random forest has 93% accuracy and SVM has 72% accuracy. Among the three models, they concluded that RF always has better performance than the other two classifiers.

John and Ammar [3] proposed a machine learning technique to the dataset collected from Kaggle. They applied feature extraction methods such as TF-IDF and sentimental analysis. Different N-gram language model-2-gram, 3-gram, and 4-gram were also applied. SVM and neural network classifiers are used for training data and their accuracy was evaluated. After evaluating the averages of accuracy, precision, recall, and f-score, the neural network has achieved an average accuracy of 91.76% and for SVM, the average accuracy was 89.87%. When they compared the proposed model with the other related work, they conclude that the proposed approach has the highest accuracy of 91.76% and an f-score of 91.9%. They found that neural network has performed better.

Shalni Prashar and Suman Bhakar [4] proposed a cyberbullying detection technique by implementing fuzzy logic and k-means clustering. Comments were classified as aggression, abuse, and threat. The proposed method results in an accuracy of 85%. Another method [8] proposed a machine learning model to prevent cyberbullying on Twitter. Data preprocessing and feature extraction were performed on the tweets fetched from twitter and that output was applied to SVM and Naive Bayes classifiers to check whether it is bullying or not. If the probability of bullying is less than 0.5, then it is not considered bullying, and if the probability is greater than 0.5, then the tweets are added to the database and also the last 10 lines of tweets from the victim's timeline are fetched. The fetched tweets are taken for the previous procedure and if that probability is less than 0.5, then that record is removed from the database. After performing these steps, they found that the accuracy of SVM is 71.25% and Naïve Bayes is 52.70%. Also, they concluded that SVM outperforms Naïve Bayes on the same dataset. In conclusion, all these approaches tried to classify comments and detect cyberbullying. Our work proposes a voting classifier that predicts the output based on the highest vote on the three classifiers. Thus we compute the capability and compare the efficiency between all the models.

## III. METHODOLOGY

### *A. Ensemble Learning*

As the name itself suggest it is a method in which more than two models are generated and ensemble to solve a particular problem. The main reason for selecting the method is because of its advantage while dealing with the lack of adequate data. Another reason is to overcome the drawbacks shown by various models. In the case of unseen data, it can't be guaranteed which model will give the correct output. So why not ensemble them, instead of choosing one, and by combining their output by, for example, majority voting can reduce the risks of underfitting as well as overfitting.

The proposed approach consists of three steps- text preprocessing, feature extraction, and classification model. Fig. 1 shows the proposed architecture.

In preprocessing step, we clean up the data by removing unnecessary text and noise. This contains tokenization, lowercasing the text, stop words removal, numbers, and, HTTPS cleaning, and lemmatize the dataset.

The next step is feature extraction. In this step, the dataset is changed to a more suitable format that can be fed into the classification model. Here we use the TF-IDF vectorization method, which extracts the features and transforms them into an array. TF-IDF gets the weight of the words with respect to the documents or sentences.

The final step is the classifier. After preparing the vector space, the classifier can train on the training corpus. After, the classifier can be used to detect cyberbullying in new chats. For the classification model, we follow an ensemble learning method in which two or more models are ensembled together to work as one.

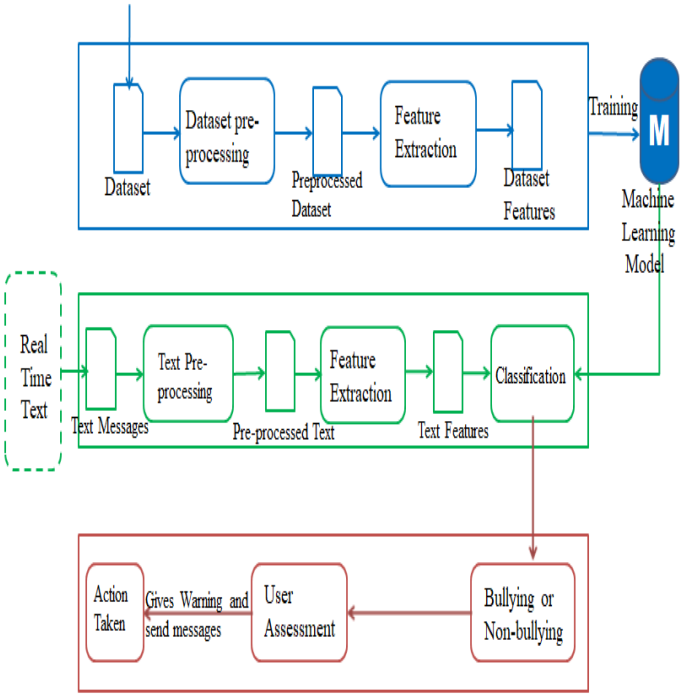


Fig.1: Proposed Architecture

### B. *Voting Classifier*

An ensemble learning model predicts an output based on the highest probability or vote gained. They operate on labels, only where

$$d_{t,j} \text{ is 0 or 1}$$

Under the condition that the classifier outputs are independent, it can be shown that the majority voting combination will always lead to a performance improvement for a sufficiently large number of classifiers. If there are a total of $T$ classifiers for a two-class problem, the ensemble decision will be correct if at least $[T/2+1]$ classifiers choose the correct class. Now assume that each classifier has a probability $P$ of making a correct decision. Then, the ensemble's probability of making a correct decision has a binomial distribution, specifically, the probability of choosing $k > [T/2+1]$ correct classifiers out of $T$ is

$$P_{ens} = \sum_{k=(T/2)+1}^{T} \binom{T}{k} P^k (1-p)^{T-k} \quad (1)$$

Then,

$$P_{ens} \longrightarrow 1, as\ T \longrightarrow \infty \text{ if } p > 0.5$$

$$P_{ens} \longrightarrow 0, as\ T \longrightarrow \infty \text{ if } p < 0.5$$

In our system, three models namely, Logistic Regression, Support Vector Machine, and Perceptron are used. Then the voting classifier ensembles all three models and gives output based on the highest majority of the vote.

Evaluation of the classifier is done using criteria like accuracy, precision, recall, and f1-score. They are calculated using the following equation:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1-score = \frac{2*precision*recall}{precision+recall} \quad (5)$$

Here TP denotes the number of true positive classes, TN denotes the number of true negative classes, FP represents the number of false-positive classes, and FN denotes the number of false negatives classes.

## IV. EXPERIMENTAL RESULTS

### A. *Dataset Description*

For this experiment, we collected the datasets from Kaggle which consists of toxic Twitter comments, YouTube comments, and comments from Form spring. The dataset contains a total of 39996 test data. Fig. 2 indicates the ratio between bullying and non-bullying comments in the dataset.

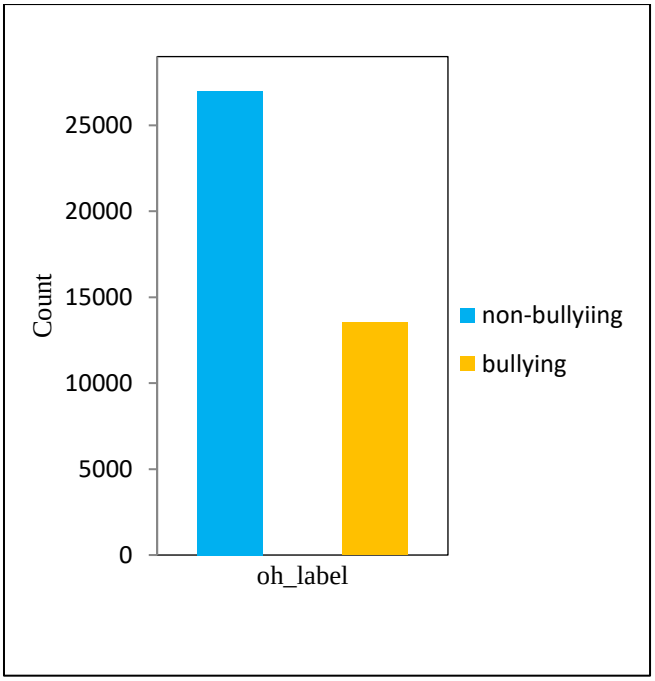


Fig. 2: Frequency of bullying and non-bullying comments in the dataset consisting of 39996 test data

### B. *Model Testing Results*

The dataset is preprocessed and then vectorized with TF-IDF and n-gram. We then split the dataset into training and testing(.8,.2) sets. Now, these datasets are fed into the three models namely Perceptron, LR, and SVM and after all these processes, the three models are ensembled into the Voting Classifier.

In this experiment, we can see that the classifier with the highest accuracy is Voting Classifier. Table I, II, III, and IV summarizes the accuracy, precision, recall, f1-score, and support of all the three models used. Fig. 3 and Fig. 4 shows the comparison of precision, recall, and, f1-score between the models on detecting non-cyberbullying and cyberbullying comments

TABLE I

PRECISION, RECALL, F1-SCORE AND SUPPORT OF PERCEPTRON NETWORK

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Label 0 | 0.91 | 0.99 | 0.95 | 6712 |
| Label 1 | 0.97 | 0.81 | 0.88 | 3287 |
| Macro avg | 0.94 | 0.90 | 0.92 | 9999 |
| Weighted avg | 0.93 | 0.93 | 0.93 | 9999 |

TABLE II

PRECISION, RECALL, F1-SCORE AND SUPPORT OF LOGISTIC REGRESSION

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Label 0 | 0.90 | 0.98 | 0.94 | 6712 |
| Label 1 | 0.96 | 0.78 | 0.86 | 3287 |
| Macro avg | 0.93 | 0.88 | 0.90 | 9999 |
| Weighted avg | 0.92 | 0.92 | 0.92 | 9999 |

TABLE III

PRECISION, RECALL, F1-SCORE AND SUPPORT OF SUPPORT VECTOR MACHINE

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Label 0 | 0.91 | 0.99 | 0.95 | 6712 |
| Label 1 | 0.97 | 0.81 | 0.88 | 3287 |
| Macro avg | 0.94 | 0.90 | 0.92 | 9999 |
| Weighted avg | 0.93 | 0.93 | 0.93 | 9999 |

TABLE IV

PRECISION, RECALL, F1-SCORE AND SUPPORT OF VOTING CLASSIFIER

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Label 0 | 0.92 | 0.99 | 0.95 | 6712 |
| Label 1 | 0.97 | 0.83 | 0.90 | 3287 |
| Macro avg | 0.95 | 0.91 | 0.93 | 9999 |
| Weighted avg | 0.94 | 0.94 | 0.94 | 9999 |

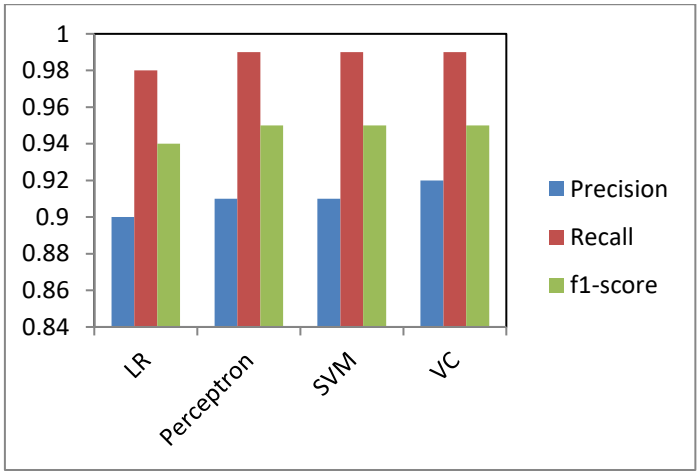


Fig. 3: Comparison of precision, recall, and f1-score between classifiers on detecting non-cyberbullying comments

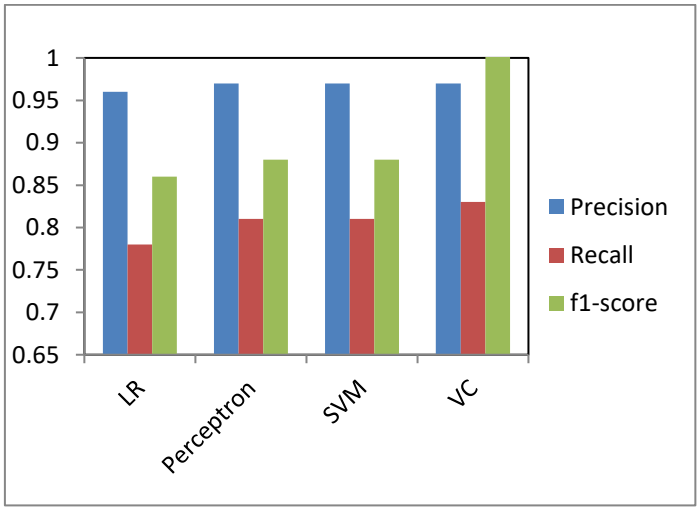


Fig. 4: Comparison of precision, recall, and f1-score between classifiers on detecting cyberbullying comments

Finally, Table V illustrates the comparison between the voting classifier with the other three classifiers in terms of accuracy. Therefore we conclude our results that the Voting Classifier has the highest accuracy and Support Vector Machine has the lowest.

TABLE V

COMPARISON OF ACCURACY BETWEEN CLASSIFIERS

| Classifiers | Perceptron | LR | SVM | VC |
|---|---|---|---|---|
| Accuracy | 0.93 | 0.92 | 0.93 | 0.94 |

### *C. Front End*

For testing our model, we created a simple public chat room with the help of Flask and SocketIO. This is to imitate the public comment section of social media. Here we blocked out any cyberbullying comments detected by our model. Fig. 5 and fig. 6 shows the user website login page and home page of a public chat window. Fig. 7 shows the user chat window. Fig. 8 illustrates how the text messages are blocked when cyberbullying comments are detected.

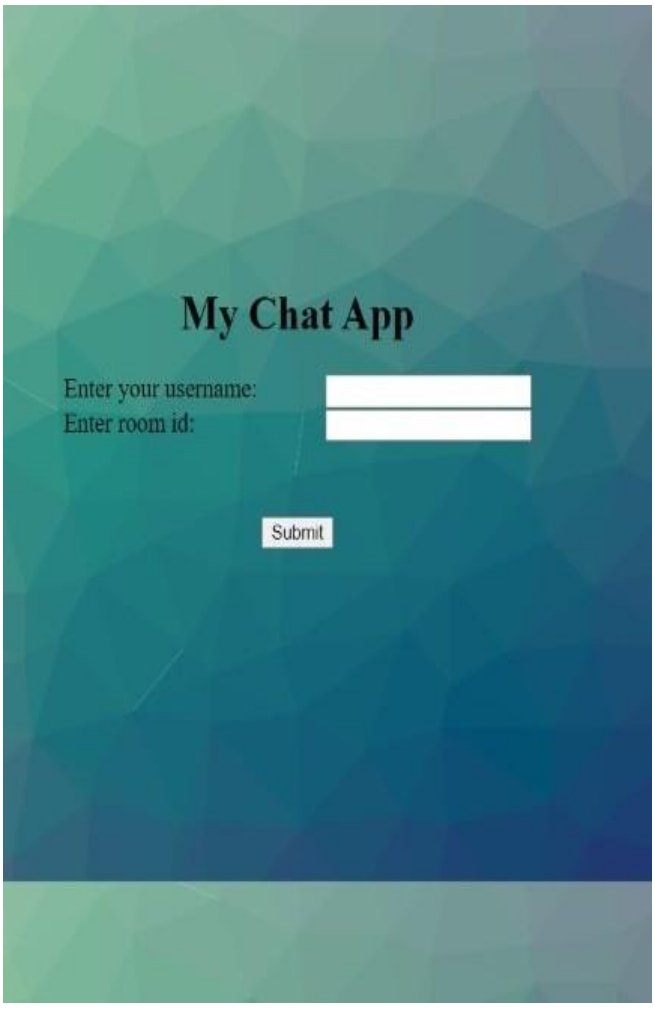


Fig. 5: User website Login page

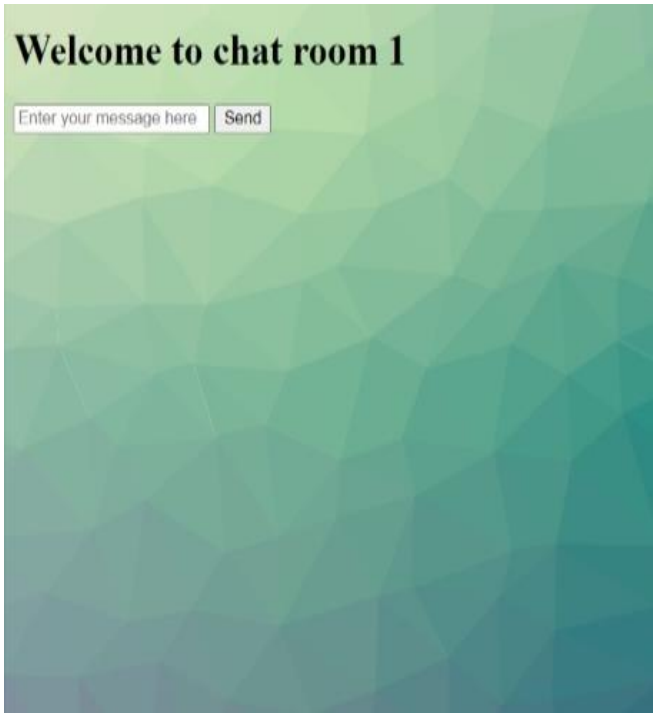


Fig. 6: Home page of user website

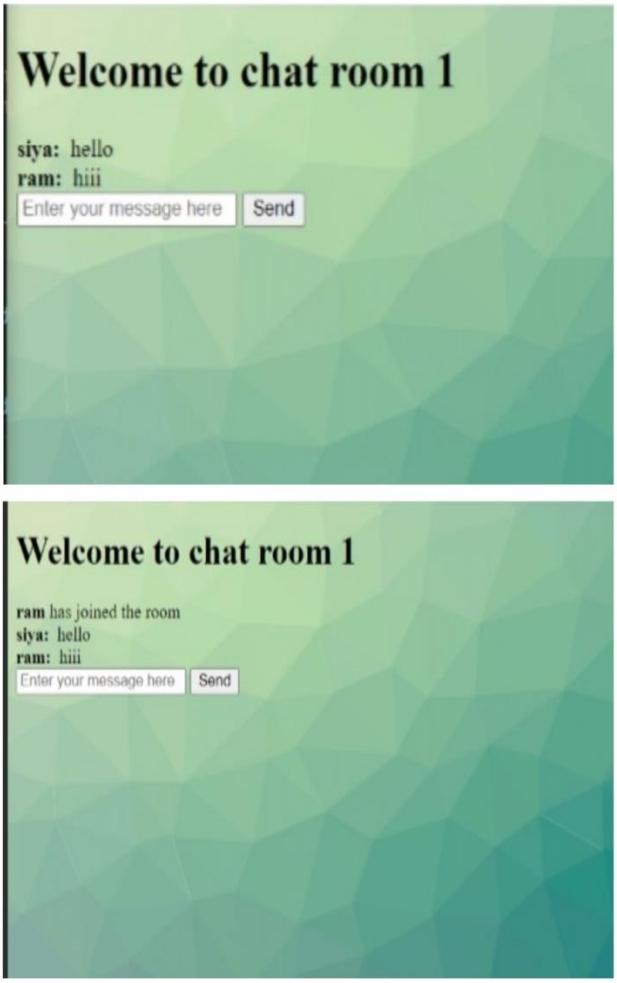


Fig. 7: User chat window


Fig. 8: Blocking comments when cyberbullying detected

## V. FUTURE SCOPE

Cyberbullying can come in many forms. We can enhance the detection of cyberbullying by combining texts with videos and images and can provide language inputs to detect sarcastic comments. So making the dataset a little more varied and including many more languages will always be a plus. Also, we can test the performance with other algorithms along with Perceptron, Logistic regression, and Support Vector Machine and compare the efficiency and accuracy in the future.

## VI. CONCLUSION

This paper aims to identify and detect cyberbullying comments using ensemble learning techniques. Here we have collected more than 30000 datasets from various sites such as Kaggle, Twitter, and YouTube. After the text pre-processing and feature extraction, we evaluated our model using Perceptron, Support Vector Machine, and Logistic Regression. Then the voting classifier is generated that ensembles all three models. We found that our voting classifier outperforms with an accuracy of 94%, which is higher than the accuracy of the other three classifiers. By acquiring this accuracy, our work will improve the real-time detection of cyberbullying comments and also help people to use social media much more safely.

## REFERENCES

[1] Bandeh Ali Talpur, Declan O'Sullivan, "Cyberbullying severity detection: A machine learning approach", Research Article: School of Computer Science and Statistics, Trinity College Dublin, Ireland, October 2020

[2] Risul Islam Rasel, Nasrin Sultana, Sharna Akhter, Phayung Meesad, "Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach", Conference: The 2nd International Conference, September 2018

[3] John Hani Mounir, Muhamed Nashat, Mostafaa Ahmed, Zeyad Emad, Ammar Mohammed, "Social Media Cyberbullying Detection using Machine Learning", Article: International Journal of Advanced Computer Science and Applications, January 2019

[4] Shalni Prashar, Suman Bhakar, "Real Time Cyberbullying Detection", International Journal of Engineering and Advanced Technology (IJEAT), Volume 9, Issue 2, December 2019

[5] Tadashi Nakano, Tatsuya Suda, Yutaka Okale, Michael John Moore, "Analysis of Cyber Aggression and Cyber-Bullying in Social Networking", Conference: 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), February 2016

[6] Giovanni Berrios, Chanhee Shin, Nishal Kallupalle, "Cyberbullying Detection System", June 2020

[7] Niraj Nirmal, Pranil Sable, Prathamesh Patil, Prof. Satish Kuchiwale, "Automated Detection of Cyberbullying Using Machine Learning", International Research Journal of Engineering and Technology (IRJET), Volume 7, Issue 2, December 2020

[8] Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, Aparna Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning", International Conference on Intelligent Computing and Control Systems (ICICCS 2020), June 2020

[9] Wanqi Li, "A Design Approach for Automated Prevention of Cyberbullying Using Language Features on Social Media", The 5th International Conference on Information Management(ICIM), May 2019

[10] Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, Aparna Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning", 4th International Conference on Intelligent Computing and Control Systems (ICICCS), June 2020

[11] Yee Jang Foong, Mourad Oussalah, "Cyberbullying System Detection and Analysis", 2017 European Intelligence and Security Informatics Conference (EISIC), December 2017

[12] Vikas S Chavan, Shylaja S S, "Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network", September 2015