Machine Learning based Approach for Detection of Cyberbullying Tweets

Rashi Shah
Student
K. J. Somaiya College of Engineering,
Mumbai-400077

Srushti Aparajit
Student
K. J. Somaiya College of Engineering,
Mumbai-400077

Riddhi Chopdekar
Student
K. J. Somaiya College of Engineering,
Mumbai-400077

Rupali Patil
Assistant Professor
K. J. Somaiya College of Engineering,
Mumbai-400077

ABSTRACT
In today's technologically sound world the use of social media is inevitable. Along with benefits of social media there are serious negative impacts as well. An important issue that needs to be addressed here is cyberbullying. An effective solution for resolving this issue is the detection of the cyber-bullying content by Machine Learning. This manuscript aims to put forward ideas regarding cyber-bullying detection on the social media platform twitter. The outcome of this manuscript is that whichever tweet is a bully tweet that is represented by the value 1, thus all the bully tweets are detected. The Twitter dataset is equally distributed into bully and non-bully tweets and fed to different machine learning models. The logistic regression classifier provides accurate classification of bully and non-bully tweets with precision of 91%, recall 94% and F1-score 93%. This work will help curb cyber-bullying, so that the users can stay at bay from victimization.

General Terms
Natural Language Processing (NLP), Machine Learning, Logistic Regression, Support Vector Machine, Random Forest Classifier, Multinomial Naïve Bayes, Stochastic Gradient Descent.

Keywords
Cyber Bullying, Machine Learning, Natural Language Processing (NLP), Twitter, Logistic Regression, Bully tweets, Non-bully tweets, Victimization, Classification.

1. INTRODUCTION
Internet era has created a powerful influence on the world it has brought people together from the bounds of their homes. Now people can have conversations and maintain relations with others without even meeting them personally. Social media has become a part and parcel of our life. There are more than billions of users using this platform as a communication gear and as a real-time data source. Social media platforms have gained immense popularity among people. Twitter is one of the most effective platforms where people of all races and ages can express themselves freely. Social media is a perfect destination for exchanging words and ideas and transmitting the best knowledge. One can get recent news at a very fast pace and at the blink of an eye. With these many boons and effort free easy technologies which have come up, there also have been ill effects of the same. Cybercriminals avail this info and use social media as a medium to commit different kinds of cybercrimes like cyber-bullying. Cyber-bullying is nothing but a form of harassment executed through digital devices. It is a world-wide problem that's growing fast. According to a recent study which was conducted by the Hindustan times, India is ranked no. 3 in cyber-bullying right behind china and Singapore. It has been recognized that because of cyber-bullying victims become dangerously timid and may get violent thoughts of revenge or even suicidal thoughts. They suffer from depression, low self-esteem and anxiety. It is worse than physical bullying because cyber bullying is "behind-the-scenes" and "24/7". Even the bully tweets or comments don't vanish, they stay for long duration and continuously affects the victim mentally. It's almost like ragging except it happens in front of thousands of mutual friends, and the scars stay forever as the messages stay forever on the internet. The hurtful and tormenting messages embarrass the victims to a level which cannot be imagined. The results are even worse and severe. In most cases, that is 9 out of 10 cases the young victims do not tell their parents or guardian out of embarrassment and get into depression or worse suicide.

The hidden scourge of cyber-bullying is something that no likes to talk about , and people have been disregarding it since a long time now, but what people don't realize is that this issue is as serious as any case of murder or other heinous crimes, the victims in these cases are mostly young and they are likely to go into depression or drop-out of school or get into alcoholism, get into drugs which ruins their entire life and indeed ruins the future of the nation because it's the youth that has the power to make the nation rise or fall.

And with such crimes which affects millions and millions of minds, progress will retard. Cyber-bullying can be of different forms: 1. Exclusion- When the victims gets deliberately excluded from any social media platform or online activities by their peers it can lead to depression. 2. Harassment- Abusive messages posted online can affect one's mental health. 3. Outing- Act of publicly insulting a person without consent. 4. Cyber-stalking- It's a dangerous type of cyber-bullying, the stalkers harass the victims through various platforms. 5. Fraping- When a user logs into your social media and impersonates them by posting inappropriate content in their name. 6 creating a fake profile. 7. Dissing- damaging someone's reputation online by posting insulting texts, photos or videos and many more.

As these issues don't happen physically cyber-bullying is not considered as a very big crime. But it should be, hence the motivation behind this project are the millions of voiceless people who are being affected. Also, its high time that we use the technologies available to us wisely and in an efficient way, this is another motivation behind this project that is to work with upcoming technologies like Natural Language Processing (NLP) & Machine Learning (ML) and resolve an important issue of the society. It is very difficult to solve all the forms of cyber-bullying, that will happen eventually, but for now our project will focus on providing a solution to the

quick detection of cyber-bullying on one of the most influential social media platforms, that is twitter and contribute in a small way in fixing this issue. This will be done by detecting the tweets which contain bully language or any kind of offensive or abusive language. Commonly used techniques are Language Processing and Text. In this paper, the aim is to classify tweets into two category bully tweets and non-bully tweets. Data preprocessing step involves TFIDF. TFIDF is nothing but a statistical technique used to review how important is a particular word in a document. It is directly proportional to number of times word appear in a doc and inversely proportional to number of docs containing that word.

Before we train the data, data needs to be cleaned as it may contain unknown symbol and errors and special characters. We address this issue by using lemmatization, stemming, omissions and elimination of certain stop words. Lemmatization is a process which eliminates the ending of words and reinstate the word to the base. Stemming is also a process of data cleaning in which it minimizes derived word to its word stem. A stop word holds no importance hence should be eliminated. The final step of data cleaning is to eliminate special and Greek characters and any foreign characters. Thus, our aim is to detect the tweets containing bullying content by developing a system to detect and classify the tweets as bully and non-bully.

2. LITERATURE REVIEW

Our work involved finding the best approach and best classifier which will accurately detect bully tweets. After reviewing two approaches i.e., lexicon-based approach and machine learning approach, the machine learning approach was chosen as it produces accurate results needed. A brief review of few major classifiers like Support Vector Classifier, Logistic Regression, Random Forest Classifier, Naive Bayes, and Stochastic Gradient Descent Classifier was done and after detailed study and testing all of these classifiers it was found that Logistic Regression has best precision and accuracy compared to others. Detailed research on the best NLP model for test analysis was done. Since our word involved cyber-bullying detection a brief review of papers related to cyber-bullying detection was done. The ever-growing use of social media has brought people together and has increased connectivity but there has also been a negative side to this. Cyber-bullying is one of the major issues among youth. Traditional studies conducted by psychologists and researchers on this matter depict more on the macroscopic level. The studies mainly focus on resolving the issue psychologically with the help of statistical data and prevention ideas. Nowadays open APIs are easily available as they are offered by prestigious social network service providers for the sake of academic research. Therefore, due to good availability of resources and effortless relevant data rather than using a limited sample of data, it is preferable to go ahead by using various methods like data crawling, data scraping which gives incentive to the event of the statistical learning of cyber-bullying which is established on Machine Learning and Natural Language Processing techniques. LDA, LSA and Bag of Words are several NLP models which are applied to detect bully tweets and outcome has been confirmed by the possibility of Automatic Detection of Cyberbullying. This idea was in accordance to one of the introductory works.

Zhao et.al [1] have presented a peculiar depiction training method for detection of cyber-bullying, entitled Embedding-enhanced Bag-of-Words. Enhanced Bag of Words progresses Bag of words features, latent semantic features, and bullying attribute in conjunction. Bullying attributes are extracted depending on word embeddings, which apprehends the semantic information trailing words when the final depiction is learned. Van Hee et. al [2] formulated a Dutch dataset of forum messages enclosing cyber-bullying and suggested and estimated a methodology for requisite implementation of data. They searched the viability of detection of bullying. Saravanaraj et.al [3] targeted on recognizing the incidence of bullying and rumors in twitter networks using type, and topic-specific classification, machine learning algorithms, and Twitter speech-act classifier, ultimately the bully tweeted and rumor spreading records will also be drawn out. And the mixing of bully recognition and rumor recognition in a single software makes recognition simpler. Cyber-bullying research often attentive on recognizing bullying ‘attacks’ and as a result neglect other or more indirect forms of cyber-bullying and post written by the bystander and victims.

Van Hee et.al [4] demonstrates a system to detect bullying automatically on social media inclusive of distinct sorts of bullying, comprising post from bullies, bystanders, and victims. The manually illustrated dataset for English and Dutch on which their system were evaluated and thereby manifested that their approach could be used for various languages. A qualitative research of results declared that false positive often include indirect bullying or insults through irony. Error rates revealed that victims are not easily recognized. Dadvar et.al advanced a gender-based bullying detection technique that used the gender attribute in improving the selective capability of classifiers, not all the users present complete record which results in a variance in the datasets it alters the efficiency of model [5].

Zhang et. al [6] proposed a novel method. An aggrandized lexicon-based technique specific to Twitter dataset was initially enforced to carry out sentiment analysis. Auxiliary dogmatic tweets could be found by performing the chi-square test on its output. Newly identified dogmatic tweets are then skilled to assign sentiment polarities and trained using binary sentiment classifier.

Training data is furnished through the lexicon-based method. Empirical experiments exhibit that the proposed approach is highly powerful and assuring. Dinakar et.al utilized Linear Discriminative Analysis to get label precise attribute and incorporate them with Bag of Words attribute to train a classifier [8]. The length of label peculiar attribute is proscribed to be less than the class numbers, which obstructs the performance hike.

Nahar et.al increased weights equivalent to bullying word by two folds [9]. They increased bullying attributes yet didn’t contemplate word semantics and the scaling performance was entirely whimsical. Apart from this, Nahar et.al [9] conjointly adopted topic models including Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) to find out topics and execute attribute selection. An alternate work is presented by Sardar Hamidian and Mona Diab which symbolizes rumor detection and classification, classification, detection and verifying is done using natural language processing tools and moreover four more aspects are checked which are date, source, location and provenance. Confirming the trueness of data is complicated [10]. Reynolds et al. [11] implemented rule-primarily based on gaining knowledge to thrive a model for detecting bullying depending on textual attributes and correlated its execution to the bag of words model (i.e., depending on the matrix of all the words that appear in the training dataset). They determined that the rule-

based method surpassed the bag-of-words model, accomplishing a recall of 78.5%.

3. METHODOLOGY

Our work involved finding the best approach and best classifier which will accurately detect bully tweets. Pre-processing of data has two steps: Collection of data and Cleaning of data. The very first and basic step is collection of data that is done in two ways. The twitter API was accessed and tweets were extracted, rest of the tweets were obtained from Kaggle dataset [17]. The dataset was divided into training and testing data. The tweets of the training data were labelled by the values 0 and 1. The bully tweets were represented by value 1 and the non-bully tweets were represented by value 0. The test data was not labelled. The next step was cleaning of the data.

3.1 Pre-processing

First Step is web crawling and data collection. We extracted few tweets from twitter. We used python's tweepy library to access Twitter API. Rests of the tweets were taken from Kaggle's dataset. The data was stored in CSV format consisting of three columns: 'id',' tweet' and 'label'. The training data had the tweets labelled and the test data was not labelled. All human languages are complex and English is one of them. A typical sentence in English consists of various verbs, nouns and is in different tenses. To find out the meaning of the sentence we need to clean our data.

Table 1: Data Dictionary

Column name	Description
Id	Serial number
Tweet	Tweets' content
Label	'0'- non-bully tweet '1' – bully tweet

Removal of special characters and numbers is also required after pre-processing. The tweets contained special characters like '@','#','&' and also numbers. To find the meaning of the sentence we removed special characters and numbers from the tweets.

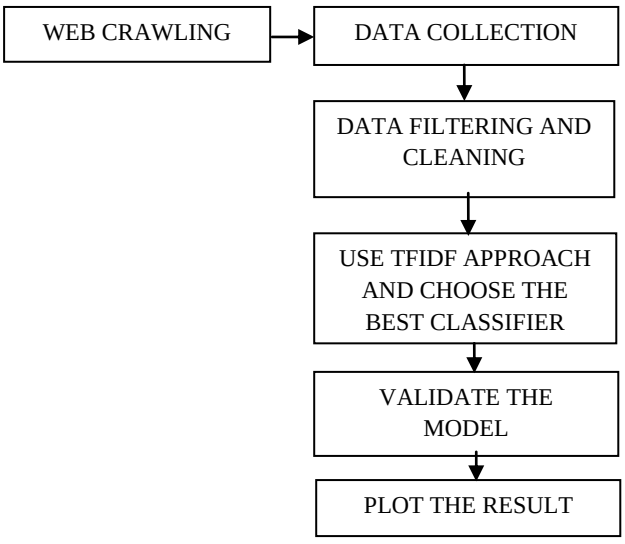

Fig 1: Flow diagram for detection of bully tweets

We did this with the help of natural language toolkit package. We used the 're' package to replace all uppercase characters to lowercase and to remove numeric values. Lemmatization is performed after removal of numbers and special characters. Lemmatization is the process of converting a word to its root form considering the context of the sentence. The module 'Wordnet Lemmatizer' which is a part of the natural language toolkit helped us lemmatizing the tweets. It has an attribute 'part of speech tag' which helps in converting the word to its root form.

For example, print (lemmatizer. lemmatizer ("advices", 'v')); #v stands for verb
>>advise
print (lemmatizer. lemmatizer ("advices", 'n')); #n stands for noun
>>advice
After cleaning the data, the next step followed is splitting the dataset. To choose the best classifier, we split the training data into two parts, one part is to train the algorithm and the other part is to check the accuracy of the algorithm. Then we apply Term Frequency-Inverse document frequency on the data.

3.2 Term frequency–inverse document frequency (TFIDF)

TFIDF is a statistical estimate of how relevant a word is in the document. It is a product of term frequency and inverse document frequency. The relevancy of the word is proportional to the number of words and is offset by number of documents that contain the word. So, words are like 'and', 'the', 'is', 'or', 'by' etc. won't be considered relevant even if their occurrence is the highest. To convert tweets into vector form, TFIDF of tweets needs to be calculated.
Term Frequency: Term Frequency is the number of occurrences of the word in the document (in our case the document is our training dataset) Inverse Document Frequency: It is calculated by total number of documents divided by the number of documents that contain that word and then taking its logarithm. So, in our case documents are our tweets. The higher the IDF score is, the rarer the word is in the document. Taking the product of TF*IDF gives the weight of the word in the document that is how relevant the word is in the document. Higher the product score, more relevant the word is. It can be mathematically expressed as:

$$tfidf(t,f,d) = tf(t,d).idf(t,d) \quad \text{[1]}$$

$$tf(t,d) = \log(1 + freq(t,d)) \quad \text{[2]}$$

$$idf(t,d) = \log\left(\frac{N}{count\ (d \in D : t \in d)}\right) \quad \text{[3]}$$

where, t=word, d=document and D= document set

Example: Consider a tweet containing 50 words where the word 'cyber-bullying' appears 2 times. The term frequency for 'cyber-bullying' is 2/50 = 0.04. Let's assume that we have 1000 tweets and the word *'cyber-bullying'* appears in 300 of these. Then, the inverse document frequency (i.e., idf) is calculated as log (1000 / 300) =0.5229. Thus, the tf-idf weight is the product of these quantities: 0.04*0.5229=0.0209. Further step is applying different classifiers. After transforming the words into numbers, these vectors are fed to classifiers such as Support Vector Classifier [13], Logistic Regression [12], Naïve Bayes' [16], Random Forest Classifier

[15], SGD Classifier [14]. The best classifier was selected by taking into consideration three factors; precision, recall and F1-score. The logistic regression classifier resulted to be the most accurate. The model was validated and the bully tweets were successfully detected.

4. RESULTS AND DISCUSSION

Total 2000 tweets are considered here for classification. Table 3 shows the equal distribution of bully and non-bully tweets from the training dataset. As per reported in literature [2,3] equal distribution of tweets gives best classification results. If the training database distribution is unequal then it may lead to wrong classification. After equal distribution, the dataset is fed to five different classifiers.

Table 2: Training dataset table

Type of tweets	No. of Tweets
Non-Bully	1000
Bully	1000
Total	2000

The best classifier among the five classifiers is identified by taking into consideration certain major factors like precision, recall, F1-score and accuracy, Specificity, MCC, Fall Out and Miss Rate values.

Table 3: Precision, Recall, F1-score and Accuracy values of various classifiers for bully tweets

Classifier	Precision	Recall	F1-score	Accuracy
SVC	0.73	0.96	0.83	0.81
Logistic Regression	**0.91**	**0.96**	**0.93**	**0.93**
Multinomial NB	0.86	0.94	0.90	0.90
Random Forest Classifier	0.98	0.73	0.84	0.86
SGD Classifier	0.90	0.95	0.93	0.92

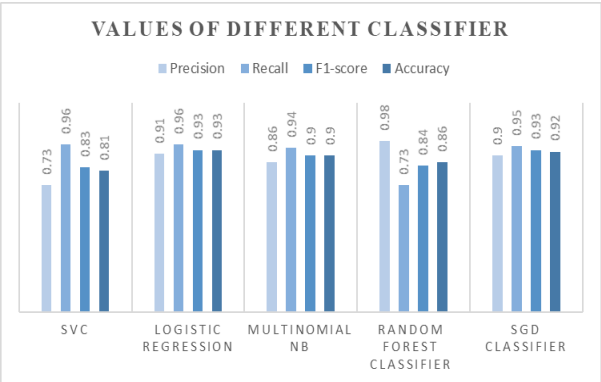

Fig 2: Precision, Recall, F1-score and Accuracy values of various classifiers for bully tweets. Logistic regression shows best values when compared with other classifiers

Table 4: Specificity, MCC, Fall Out, Miss Rate and Mean Square Error values of various classifiers for bully tweets

Classifier	Specificity	MCC	Fall Out	Miss Rate	MSE
SVC	0.65	0.65	0.34	0.03	0.512
Logistic Regression	**0.90**	**0.87**	**0.09**	**0.03**	**0.066**
Multinomial NB	0.85	0.80	0.15	0.06	0.1
Random Forest Classifier	0.98	0.74	0.01	0.27	0.166
SGD Classifier	0.90	0.84	0.10	0.05	0.07

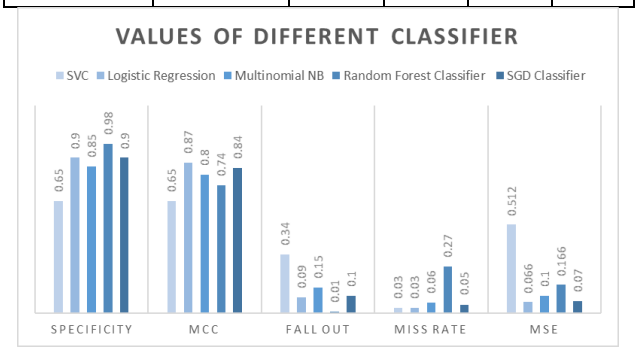

Fig 3: Specificity, MCC, Fall out, Miss rate and Mean Square Error values of various classifiers for bully tweets. Logistic regression shows the desired values when compared with other classifiers

Precision determines the quantity or proportion of positive identifications which are 100% correct. Precision is given by:

$$Precision = \frac{TP}{TP + FP} \quad \text{[4]}$$

where TP-True Positive & FP-False Positive. A model that does not produce a single false positive is absolutely accurate and has a precision of 1.0. Closer the precision to 1 better the classifier. The precision for bully tweets for the logistic regression classifier is, 0.91 and it is better compared to other classifiers. (as shown in Fig. 2 & Table 3).

Recall determines the actual proportion of positives which are identified correctly. Mathematically, recall is given by the formula:

$$Recall = \frac{TP}{TP + FN} \quad \text{[5]}$$

where FN-False negative. A model with recall 1.0 produces no false negatives. As shown in fig. 10 we can see that the recall value of bully tweets is 0.94, the recall value of bully tweets of other classifiers is low, because they have a greater number of false negatives for bully tweets, hence logistic regression has more accuracy comparatively.

The F1 score conveys the balance between the precision and the recall. The formula for F1-score is given below:

$$F1\ Score = 2 * P * \frac{R}{P+R} \quad \text{[6]}$$

where P is precision and R is recall. It can also be referred to as the F1 Score or the F1 Measure. The F1 score of the logistic regression classifier of bully and non-bully tweets is 0.93, which is better than the other classifiers.

Accuracy is the determination of how well a binary classification test accurately defines or removes a condition. It is a test parameter to measure binary precision, as shown in the table 3 and figure 2, the Logistic regression classifier provides the best accuracy of 93%, the formula is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{[7]}$$

It is the proportion of accurate predictions (both true positive and true negative) among the total number of cases examined.[18] It is often referred to as the" Rand accuracy "or" Rand index ". The Specificity of the test, is nothing but the true negative rate (TNR), which is the proportion of negative samples testing negative.

$$Specificity = No.of\ True\ Negative\ (TN) \quad \text{[8]}$$

Logistic regression (LR) gives the best Specificity i.e., 0.90 as shown in the table 4 and figure 3. The Matthews correlation coefficient (MCC) or phi coefficient is used as a measure of binary (two-class) classification quality.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad \text{[9]}$$

The coefficient takes true positives, false positives, true negatives and false negatives into account and is generally considered a balanced measure that can be used even if the classes are of very different sizes. In essence, the MCC is a correlation coefficient between the binary classifications observed and predicted; it returns a value between -1 and +1. A coefficient of +1 is a perfect prediction, 0 is no better than a random prediction, and −1 shows a total discrepancy between prediction and observation.

As shown in the table 4 and figure 3 MCC for LR is 0.87, which is better compared to other classifiers.

Miss rate is the false negative rate. The formulae for miss rate are:

- $$Miss\ Rate = \frac{FN}{FN + TPR} \quad \text{[10]}$$
- $$Miss\ Rate = 1 - TPR \quad \text{[11]}$$

where TPR=True Positive Rate.
Lower the percentage of miss rate better is the classification.

As shown in the table 4 and figure 3 the miss rate for LR is 0.03, which is lesser compared to other classifiers.

Fall out is the false positive rate. The formulae for fall out are

- $$Fall\ Out = \frac{FP}{FP + TNR} \quad \text{[12]}$$
- $$Fall\ Out = 1 - TNR \quad \text{[13]}$$

where TNR=True Negative Rate. Lower the percentage of fall out, better is the classification.

As shown in the table 4 and figure 3 the Fall out for LR is 0.09, which is lesser compared to other classifiers.

The Mean Squared Error (MSE) is the average squared difference between the desired values and actual value. It is a risk function, corresponding to the expected value of the squared error loss. It is always non – negative and values close to zero are efficient. The MSE is the second moment of the error (about the origin) and thus involves both, the variance of the estimator and its bias.

$$Mean\ Squared\ Error = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2 \quad \text{[14]}$$

where y_j = original y values

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j + \hat{\epsilon}_j \quad \text{[15]}$$

Equation 15 is the equation of Regression Line. As shown in the table 4 and figure 3 the Mean Squared Error for LR is 0.066, which is lesser compared to other classifiers. From Table 4, comparing our algorithm with other algorithms we can conclude that machine learning and tf-idf approach for text classification gives better results than using Lexicon and bag of words approach. Zhao et. al [1] and Van Hee et. al [4] have used Bag of words approach and Zhang et. al [6] has used Lexicon based approach. Comparing the F1 scores of all the four algorithms we can see that our approach has the highest F1 score. One disadvantage of machine learning approach is that it needs more data than other approaches and for supervised machine learning approach it needs labeled data.

Table 4: Comparison with other existing system

	F1 score (%)	**Precision (%)**	**Recall (%)**
Zhao et. al [1]	75.6	77.8	76.6
Van Hee et. al [4]	64.26	73.32	57.19
Zhang et. al [6]	74.9	68.7	82.7
Our algorithm	**93.0**	**91.0**	**96.0**

5. CONCLUSIONS

The twitter data analysis was performed with the aim of detecting the bully tweets. In this manuscript, the detailed procedure of Cyber-bullying detection on twitter platform is explained, right from data collection, separation into training and testing data, selecting the best classifier, validating the model. The selection of the best classifier was done by plotting the model's precision, F1-score, recall, accuracy,

Specificity, MCC, Fall Out and Miss Rate values. From the results obtained it can be concluded that the logistic regression classifier is the most accurate among all the other classifiers which has 91% precision, 96% recall and 93% F1 score, 93% accuracy, 90% Specificity, 87% MCC, 9% Fall Out and 3% Miss Rate. The future scope of this work involves some ideas or further steps that can be taken to root out cyber-bullying from twitter completely. These are the ideas that can be implemented in the future: Once the bully tweets are detected the person who has received the tweet will be given three options that is 1. report the sender, 2. delete the tweet, 3. both report and delete generate. Another step that can be taken is to automatically report if bullying content is detected, so that those tweets or comments get deleted and it doesn't affect the user. In terms of punishment for the bullies who have been reported by the victims more than once will be blacklisted and a warning will be given. If the bullying continues further then strict action will be taken against that person.

6. REFERENCES

[1] Zhao, R., Zhou, A. and Mao, K., 2016, January. Automatic detection of cyber-bullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking* (pp. 1-6).

[2] Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W. and Hoste, V., 2015. Automatic detection and prevention of cyber-bullying. In *International Conference on Human and Social Analytics (HUSO 2015)* (pp. 13-18). IARIA.

[3] Saravanaraj, A., Sheeba, J.I. and Devaneyan, S.P., 2016. Automatic detection of cyber-bullying from twitter. *Int. J. Comput. Sci. Info. Technol. Secur*, *6*.

[4] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W. and Hoste, V., 2018. Automatic detection of cyber-bullying in social media text. *PloS one*, *13*(10).

[5] Dadvar, M., Trieschnigg, D. and de Jong, F., 2014, May. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence* (pp. 275-281). Springer, Cham.

[6] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M. and Liu, B., 2011. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, *89*.

[7] Amolik, A., Jivane, N., Bhandari, M. and Venkatesan, M., 2016. Twitter sentiment analysis of movie reviews using machine learning techniques. *international Journal of Engineering and Technology*, *7*(6), pp.1-7.

[8] Dinakar, K., Reichart, R. and Lieberman, H., 2011, July. Modeling the detection of textual cyber-bullying. In *fifth international AAAI conference on weblogs and social media*.

[9] Nahar, V., Li, X. and Pang, C., 2013. An effective approach for cyber-bullying detection. *Communications in Information Science and Management Engineering*, *3*(5), p.238

[10] Sardar Hamidian and Mona Diab. Rumor Detection and Classification for Twitter Data, IARIA (2015), 71-77, SOTICS2015: The Fifth International Conference on Social Media Technologies, Communication, and Informatics.ISBN:978-1-61208-443-5.

[11] K. Reynolds, A. Kontostathis, and L. Edwards, "December 18–21, 2011, Honolulu, Hawaii. IEEE Computer Society, Dec. 2011, pp. 241–244, IEEE, ISBN: 978-0-7695-4607-0,

[12] https://en.wikipedia.org/wiki/Logistic_regression#:~:text=Logistic%20regression%20is%20a%20statistical,a%20form%20of%20binary%20regression

[13] https://en.wikipedia.org/wiki/Supportvectormachine

[14] https://towardsdatascience.com/how-to-make-sgd-classifier-perform-as-well-as-logistic-regression-using-parfit-cc10bca2d3c4

[15] https://en.wikipedia.org/wiki/Random_forest

[16] https://scikitlearn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#:~:text=The%20multinomial%20Naive%20Bayes%20classifier,tf%2Didf%20may%20also%20work

[17] https://www.kaggle.com/vkrahul/twitter-hate-speech

[18] https://www.jmir.org/