International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 8 Aug. 2022, pp: 1172-1184 www.ijaem.net ISSN: 2395-5252

Cyber Bullying Detection in Twitter using MachineLearning Algorithms

Hadiya E M
Department of Computer ScienceCUSAT
Kerala, India

Submitted: 10-08-2022 Revised: 20-08-2022 Accepted: 22-08-2022

ABSTRACT—With widespread usage of online social networks and its popularity, social networking platforms have given us incalculable opportunities than ever before, and its benefits are undeniable. Despite benefits, people are humiliated, insulted, bul- lied, and harassed by anonymous users, strangers, or peers.Cyberbullying is defined as “willful and repeated harm inflicted through the use of electronic devices” [10]. In this study, I have proposed a cyber bullying detection framework to generate features from dataset of Twitter content by extracting different kinds of features.I applied Semantic, Sentiment,Syntactic and Pragmatic features along with the conventional feature extraction methods like TFIDF and CountVectorizer. Extracted features were applied with Multi Layer Perceptron,Logistic Regression, Naive Bayes, KNN, Random Forest, and Support Vector Machine algorithms. Results from the experiments are promising withrespect to classifier accuracy and F-measure metrics.I compared the results of various machine learning algorithms with different feature combinations. Findings of the comparison indicate the significance of the proposed features in cyber bullying detection.

Index Terms—Cyber bullying,features, TFIDF,Count Vector- izer,Machine learning Algorithms

I. INTRODUCTION

With the spread of mobile technologies, cyber bullying has become an increasing problem, especially among teenagers. Atthe appropriate time, cyberbullying comes in different various structures. It does not really mean hacking somebody’s profile or presenting to be another person. It likewise incorporates posting negative remarks about someone or spreading bits of hearsay to criticize somebody. According to recent studies almost 43% of teenagers in the U.S. revealed to be victimsof cyber bullying. Even though the problem is now heavily considered from a social point of view, computational studies in this field are largely yet unexplored and only few researches on cyber bullying are available.

II. MOTIVATION

Among the numerous existing social networks, Twitter isa critical platform and a vital data source for researchers.It is the most popular public blogging network operating in real-time, in which news often appears before it appears in official sources. Characterized by its short message limit (now 280 characters) and unfiltered feed, Twitter use has rapidly increased, with an average of 500 million tweets posted daily, particularly during events [11].However with Twitter becoming a notable and an actual communication channel, a study has reported that Twitter is a “cyberbullying playground”. For this reason, data crawled from Twitter was considered as a good source for our cyberbullying research.

Cyberbullying may negatively impact the victim’s self-esteem, academic achievement and emotional well-being.The self- reported [8] effects of cyberbullying include negative effects on school grades and feelings of sadness, anger, fear, and depression. In extreme cases, cyber-bullying could even lead to self-harm and suicidal thoughts. These findings demonstrate that cyberbullying is a serious problem, the consequences of which can be dramatic and tragic.Attempts for the early detection of cyberbullying is therefore is of key importance to youngsters mental well-being.

III. PROBLEM STATEMENT

The objective of this project is to develop a model for cyber bulling detection based on machine learning approachesusing Twitter datasets available from public sources. Differ- ent features like semantic,sentiment,syntactic and pragmatic are extracted from the tweets and trained using various machine learning algorithms including SVM, KNN,Random Forest,Naive Bayes, MLP and Logistic Regression. This study gives the idea of how important those features are for cyber bullying detection in tweets.

IV. LITERATURE SURVEY

In the paper 'Detecting A Twitter Cyber bullying Using Machine Learning' done by Rahul Ramesh Dalvi,Sudhanshu Baliram Chavan,Aparna Halbe [12], a machine learning modelis proposed to detect and prevent bullying on Twitter.Two classifiers - SVM and Naive Bayes are used for training and testing the social media bullying content.Both Naive Bayes and SVM (Support Vector Machine) were able to detect the true positives with 71.25% and 52.70% accuracy,respectively.It shows that SVM outperforms Naive Bayes of similar work on the same data set.

Maral Dadvar and Franciska de Jong [4] proposed that the incorporation of the users' information, their characteristics, and post-harassing behaviour, alongside the content of their conversations, will improve the accuracy of cyberbullying detection. They investigated the cyberbullying detection from two perspectives. First, which is the conventional way, the users' behaviour will be considered only in one environment, for instance, the user's comments on a video on YouTube.They envisioned an algorithm that would go through the comments' text and would classify them as either bullying or non-bullying. At this phase of the experiment, they hypothesized that including the users' characteristics – either the bully or the victim - such as age and gender, will improve thedetection accuracy.They have investigated the gender-based approach for cyberbullying detection in MySpace, in which improvements are observed in classification. Their analysis showed that author information can be leveraged to improve the detection of misbehaviour in online social networks.

In the paper done by Michele Di Capua,Emanuel Di Nardo, Alfredo Petrosino [1],they developed a model of cyber-bulling aggression, based on a hybrid set of features, starting with classical textual features but also based on the so-called "social features".These features are related to social behavior and their peculiarities are strictly related to the social platform ana- lyzed.They used data from formspring which got the accuracy of 73% and for the YouTube dataset ,acquired the accuracy of 69%.

Reynolds, Kontostathis, and Edwards [5] conducted a study on cyber aggression detection using two different methods:rule-based learning and bag-of-words model, using data extracted from Formspring.me, a social media to ask 1- on-1 questions. The rules defined consisted of presence of bad words and anonymity, because they argued that anonymity might promote user's tendency to bully or harass. Although this study successfully detect cyberbullying, the high number of false positive was an issue.

Chen [3] had proposed a technique to detect offensive language constructs from social networks through the analysis of features that are related to the users writing styles, struc- tures,and certain cyberbullying contents to identify potential bullies.The basic technique used in this study is a lexical syntactic feature that was successfully able to detect offensive contents from texts sent by bullies. Their results indicated a very high precision rate 98.24%, and recall of 94.34%.

Amanpreet Singh et al. [15] has reviewed many previous research papers related to machine learning models, prepro- cessing techniques, evaluation of machine learning models,etc. They have discussed used methodology, datasets, conclu- sions/findings, content-based features, demerits, technique andused models, preprocessing steps used for the model. Forresearching purposes, they have explored Scopus and the IEEE Xplore virtual library, ACM Digital Library. Using citations, 51 academic papers were discovered. Based on concluding arguments, abstracts, and titles, 18 papers were found notto apply to the survey so 18 papers were discarded. In this paper for the survey, they have reviewed 27 papers from 33 papers after filtration. In each of the 27 research papers, binary classification is used for cyberbullying detection. And most of them have used the Support Vector Machine (SVM) algorithm for detection.

VandanaNandakumar et al. has done another survey on Twitter data using Naive Bayes classifier algorithm and Sup- port Vector Machine model [9]. The feature probabilities are calculated using Naive Bayes Classifier Algorithm. A graph is plotted comparing among the two algorithms, Naive Bayes Classifier Algorithm and Support Vector Machine. Compar- ison on the basis of precision factor is also done with thefact that the probabilities for each feature set are calculated independently from the twitter dataset and can evaluate the performance by predicting the output variable. The plotted graph shows that Naive Bayes classifier shows better precision than support vector machine model. They concluded that, for text data classification, Naive Bayes classifier shows better performance than the SVM model.

J. Yadav, et al. [18] proposes a new approach to cyber- bullying detection in social media platforms by using the BERT model with a single linear neural network layer on top as a classifier. The model is trained and evaluated on the Formspring forum and Wikipedia dataset. The proposed model gave a performance accuracy of 98% for the Formspring dataset and of 96% for the Wikipedia dataset which is relatively high from the previously used models. The proposed model gave better results for the Wikipedia dataset due to its large size and without the need for oversampling whereas the Form spring dataset needed oversampling.

Trana R.E., et al. [16] goal was to design a machine learning model to minimize special events involving text extracted from image memes. The author has compiled a database containing approximately 19,000 text views published on YouTube. This study discusses the effectiveness of the three machine learning machines, the Uninformed Bayes, the Sup- port Vector Machine, and the convolutional neural network used on the YouTube database, and compares the results with the existing Form databases. The authors further investigated algorithms for Internet cyberbullying in sub-categories within the YouTube database. Naive Bayes surpassed SVM and CNN in the following four categories: race, ethnicity, politics, and generalism. SVM has passed well with the inexperienced Naive Bayes and CNN in the same gender group, and allthree algorithms have shown equal performance with central body group accuracy.The results of this study provided data that can be used to distinguish between incidents of abuse and non-violence. Future work could focus on the creation of a two-part segregation scheme used to test the text extracted from images to see if the YouTube database provides a better context for aggression-related clusters.

N. Tsapatsoulis, et al. [17] presented a detailed review of cyberbullying on Twitter. The importance of identifying different abusers on Twitter is given. In the paper, various practical steps required for the development of an effective and efficient application for cyberbullying detection are de- scribed thoroughly. The trends involved in the categorization and labeling of data platforms, machine learning models and feature types, and case studies that made use of such tools are explained. This paper will serve as an initial step for the project in Cyberbullying Detection using Machine learning.

G. A. León-Paredes et al. [7] have explained the devel- opment of a cyberbullying detection model using Natural Language Processing (NLP) and Machine Learning (ML). A Spanish cyberbullying Prevention System (SPC) was devel- oped by applying machine learning techniques Naive Bayes, Support Vector Machine, and Logistic Regression. The dataset used for this research was extracted from Twitter. The max- imum accuracy of 93% was achieved with the help of three techniques used. The cases of cyberbullying detected with the help of this system presented an accuracy of 80% to 91%on average. Stemming and lemmatization techniques in NLP can be implemented to further increase the accuracy of the system. Such a model can also be implemented for detection in English and local languages if possible.

Rasel, Risul Islam, et al. [13] focuses on the removal of the comments made on social networks, and the analysis of the question as to whether these observations provide an offensive meaning. The reactions can be divided into three categories: offensive, hate speech, and neither of the two. The proposed model classifies the notes on the species, with an accuracy of more than 93%. Latent Semantic Analysis (LSA) has been used as a feature selection method to reduce the size of the input data. In addition to standard feature extraction methods such as tokenization and N-gram, TFIDF was applied to detect the important notes. They made three different machine learning models, Random Forest, Logistic Regression, and Support Vector Machines (SVMs) to perform the calculation, analysis, forecasting, and a teasing comment.

In the paper by Chatzakou, Despoina and Kourtellis [2], they discussed the creation of a quality tweet corpus related to harassment and annotated that with respect to the five types of harassment content (i) sexual, (ii) racial, (iii) appearance related, (iv) intellectual, and (v) political. This is the first corpus that takes content type into account. Furthermore, they have also developed a lexicon of content-specific offensive words along with a generic category of offensive words. They first crawled data from Twitter using this content-tailored offensive lexicon. As mere presence of an offensive word is not a reliable indicator of harassment, human judges annotated tweets for the presence of harassment. Their corpus consists of more than 20,000 annotated tweets for the five types of harassment content and is available on the Git repository.They also made this dataset available to encourage comparative analysis of harassment detection algorithms.

V. PROPOSED SYSTEM

I propose a possible solution for automatic detection of the bully traces, especially twitter posts containing harmful text or sentence that could possibly lead to a cyber bullying episode. I shall show that using both techniques derived from Natural Language Processing and also on the basis of Senti- ment, Semantic, Syntactic and Pragmatic Analysis approach, considering, as an assumption, that a cyber bullying post is an extremely negative message, in the preprocessing data stage, and the subsequent adoption of unsupervised machine learning algorithms, for the detection phase, can lead to reliable results.

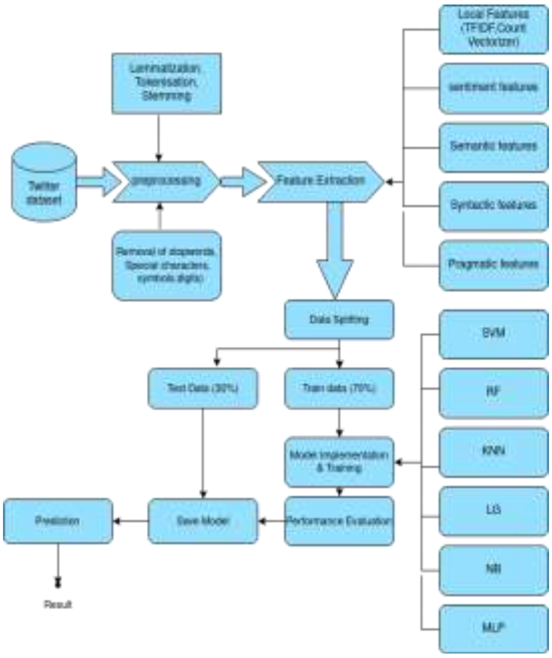

Fig. 1. Proposed Model Framework

VI. METHODOLOGY

In this section,the cyberbullying detection framework is described which consists of two major parts as shown in Figure1.

The first part includes NLP (Natural Language Processing) and the second part, ML (Machine learning). In the first phase, Twitter datasets containing bullying texts, messages or posts or tweets are collected and preprocessed by using different feature extraction methods. Then they are used to train the machine learning algorithms for detecting any harassing or bullying message.

A. Preprocessing

The preprocessing of data includes the removal of symbols, special characters, digits, short words, and stop words from each post in the dataset. Tokenization and stemming are also done as part of preprocessing.

1) Natural Language processing: The real world posts or texts contain various unnecessary characters or text. For example, numbers or punctuation are irrelevant to bullying detection. Before applying the machine learning algorithms to the tweets, we need to clean and prepare them for the detection phase. Various processing tasks including removal of all irrelevant characters like stop words, punctuation and numbers, tokenizations, stemming etc.Tokenization is the pro-cess of breaking a text corpus up into most commonly words, phrases, or other meaningful elements, which are then called tokens.word tokenize method of nltk module is applied for tokenization.Stop-words and stemming procedures are alsoperformed before feature extraction. Stop words are defined as the insignificant words that appear in document which are not specific or discriminatory to the different classes. Stemming refers to the process of reducing words to their stems or roots. For instance, singular, plural and different tenses are consolidated into a single word. We used WordNetLemmatizerfrom nltk for this process. After the preprocessing, we are going to extract the important features from the tweets.

B. Feature Extraction

We developed a model based on a hybrid set of features, starting with classical textual features like TFIDF and Count Vectorizer but also based on the Syntactic,Sentiment [14], Semantic, Pragmatic features. This model avoided a bag- of-words (BoW) approach because this approach does notconsider the position of words in a sentence and also because in the BoW model the feature space can be significantly large. In order to accomplish our task, we manually build some features considering the cyber bullying problem from differentpoints of view then divided the features in groups, to distin- guish them based on pure text analysis. The distinct features group, is divided into:

- Local Features
- Syntactic features
- Semantic features
- Sentiment features

- Pragmatic features

1) Local Features: Local features are described as features extracted from the post itself, assessed using TFIDF (Term Frequency–Inverse Document Frequency) rule and also Count Vectorizer.We found that utilizing combined features resulted in better performance compared to elementary TFIDF.

- This is one of the first features that we consider for our model. TFIDF (Term Frequency-Inverse Document Frequency) is a statistical measure that can evaluate how relevant a word is to a document in a collection of documents. In bag of words, every word is given equal importance while in TFIDF, the words that occur more frequently should be given more importance as they are more useful for classification.
- Count Vectorizer is a method to convert text to numerical data. To show you how it works, let's take an example: text = 'Hello my name is James', 'this is my python notebook'

This text will be transformed to a sparse matrix. We have 8 unique words in the text and hence 8 different columns each representing a unique word in the matrix. The row represents the word count. Since the words 'is' and 'my' were repeated twice we have the count for those particular words as 2 and 1 for the rest. Count Vectorizer makes it easy for text data to be used directly in machine learning and deep learning models such as text classification.

2) Syntactic Features: These features are generally ob- tained by statistical analysis of documents (tweets):

- **Presence of bad or abusive or profanity words**: From literature, it is quite evident and intuitive that some "bad" words make a text a suitable candidate to be labeled as a possible cyber bullying sentence. As done in other works,we have identified a list of insults and swear words (835 terms), which is collected from publicly available online sources.
- **Bad or abusive or profanity words count**: In our model, we also stored the count of "bad" words as a single feature. This feature is equivalent to the number of bad words that appear in each tweet.

3) Semantic Features: These are the ones that describehow an internet user uses punctuation, capitalized words, and interjections, etc. Although hate speech on social networks andmicro blogging websites do not have a specific and a commonuse of punctuation or employment of capitalization, in some cases, some of these reflect some sort of segregation or others, such as the following example:

"Why don't you simply go back to YOUR COUNTRY and leave us in peace?"

The tweet is obviously offensive and shows some hate, however, there is no explicit use of hate words, or any sentimental word (except the word "peace" which is obvi- ously a positive word). Therefore, we believe that punctuation features, including the capitalization, the existence of question and exclamation marks, etc. help in detecting hateful speech, and they cannot be simply discarded.

- **Density of upper case letters**: This feature is based on Dadvar et al. [4] results. The presence of capital letters in a text message is selected as a feature, considering it as possible 'shouting' at someone behavior, as commonly treated in social networks netiquette. This feature is given by the the number of upper case letter in each tweet.
- Just like capital letters, **exclamations ,full stops, quotes and questions marks** can be considered as important in comments. It can be connected to the strong (usually bad) language.We consider identifying the presence of exclamation points and question marks as a feature in our model will be very helpful.

4) Sentiment Features: Sentiment analysis and cyber bully- ing detection were strictly correlated topics. In a cyber bullyingpost, there is a wide range of emotions that can be used to identify victims. Hence it still makes sense to use sentiment- based feature as the most basic feature that allow the detection of hate speech. This is because hate speech is most likelyto be present in a "negative" tweet, rather than a "positive" one. Consequently, we first extract features that would help to determine whether a tweet is positive, negative or neutral. As mentioned above, the detection of the polarity in itself is not the purpose of this work, but an extra step to facilitate the main task, which is the detection of hate speech. Therefore, from each tweet we extract the following features:

- The total score of positive words (PW),
- The total score of negative words (NW),
- The total score of neutral words (NW).

5) Pragmatic features: In this modern world, Emojis are essential to communicate emotion, something that words can-not portray. Emotional signals are any information that could be correlated with sentiment polarity of a sentence. Recentlyin social media, users adopt visual cues that are strongly associated with their emotional states. These cues, known as emoticons (or facial expressions), are widely used to show the emotion that a user's post represents. We have used emoji module of python package to identify the presence of emojis in the tweets. There are a total of 1853 emojis in the module.

C. Machine Learning Algorithms

This module involves in applying various machine learning approaches like Random Forest, Support Vector Machine, Naive Bayes ,K-Nearest Neighbour, Multi-Layer Perceptron, Logistic Regression to detect the bullying message and text. The classifier with the highest accuracy is discovered for our two public cyberbullying twitter dataset.

1) Support Vector Machine: Support Vector Machine (SVM) is a supervised machine learning algorithm which can be applied in both classification and regression like a decision tree. It can distinguish the classes uniquely in n-dimensional space. Thus, SVM produces a more accurate result than other algorithms in less time. In practice, SVM constructs set a of hyper planes in a infinite-dimensional space and is implemented with kernel which transforms an input data space into the required form.

2) Random Forest: Random Forest classifier consists of multiple decision tree classifiers. Each tree gives a class prediction individually. The maximum number of the pre-dicted class is our final result. This classifier is a supervised learning model which provides accurate result because several decision trees are merged to make the outcome. Instead of relying on one decision tree, the random forest takes the prediction from each generated tree and based on the majority votes of predictions, and it decides the final output. For example, if we have two classes namely A and B and the most of the decision tree predict the class label B of any instance, then RF will decides the class label B as follows:f(x) = majority vote of all tree as B

3) Naive Bayes: Naive Bayes is an efficient machine learning algorithm based on Bayes theorem. The algorithm predicts depending on the probability of an object. The binary and multi-class classification problems can be quickly solved using this technique. Based on Bayes Theorem ,it finds the probability of an event occurring given the probability of another event that has already occurred as follows:

$$\frac{p(\mathbf{X}|y) * p(y)}{\mathbf{X}}$$

Here, where, the class variable is denoted by y and X is a dependent feature vector of length n as $\mathbf{X} = x_1, x_2, x_3 \ldots x_n$

4) K-Nearest Neighbour: KNN is one of the simplest Machine Learning algorithms based on Supervised Learning technique.It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

5) Logistic Regression: LG is one of the most popular Ma- chine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent vari- ables.It uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.

6) Multi Layer Perceptron: MLP is a supplement of feed forward neural network. It consists of three types of lay- ers—the input layer, output layer and hidden layer.The input layer receives the input signal to be processed. The required task such as prediction and classification is performed by the output layer. An arbitrary number of hidden layers that are placed in between the input and output layer are the true computational engine of the MLP. Similar to a feed forward network, in this, the data flows in the forward direction from input to output layer. The neurons in the MLP are trained with the back propagation learning algorithm. They are designed to approximate any continuous function and can solve problems which are not linearly separable. The major use cases are pattern classification, recognition, prediction andapproximation.

VII. EXPERIMENTS AND RESULTS

A. Datasets

In our study, we have used 3 annotated datasets with 1065 tweets, 9057 tweets and 22890 tweets. The reasons for selecting this dataset include:

1) It is publicly available on Git repository (https://www.kaggle.com/datasets/dataturks/da

taset-for-detection-of-cybertrolls)and from https://data.mendeley.com/datasets/jf4pzyvnpj/1

2) It is well-suited for our study.

The texts or comments were classified into two types asfollows:

- **Non-bullying Text** This type of comments or posts or tweets are non-bullying or positive tweets. For example, the tweet like ”This photo is very beautiful” is positive and non-bullying tweet.
- **Bullying Text** This type belongs to bully type comments or harassment. For example, ”go away bitch” is a bullying text or comment and we consider as negative tweet.

Initially the third dataset contained only 20001 tweets and imbalanced being non-bullying tweets as the majority. So we used the dataset of 1065 tweets and tweets from other sources to be merged manually into this dataset so as to solve the imbalance, finally resulting in a dataset of 22890 tweets.

B. Results

We have used six machine learning algorithms namely,Support Vector Machines (SVM), Naive Bayes (NB), Logistic Regression(LG), K-Nearest Neighbor(KNN),Multi- Layer Perceptron(MLP) and Random Forest (RF) to classify tweets as bullying or non-bullying. In this section, we first describe the datasets for the experiment and then discuss about the results.

TABLE I
TWITTER DATASETS USED

Number of tweets	**Non Bullying** 0	**Bullying** 1
dataset with 9057 tweets	4852	4204
dataset with 22890 tweets	12179	10711

TABLE II
ACCURACY ACHIEVED BY APPLYING EACH FEATURES ALONE

Models *A*	**TFIDF** *B*	**Semantic** *C*	**Sentiment** *D*	**Count Vector-izer** *E*	**Pragmatic-Syntactic** *F*
SV M	73	74	77	77	75
RF	85	82	93	79	75
LG	73	74	77	75	75
KNN	55	74	83	51	75
NB	44	37	74	43	74
MLP	79	74	77	80	75

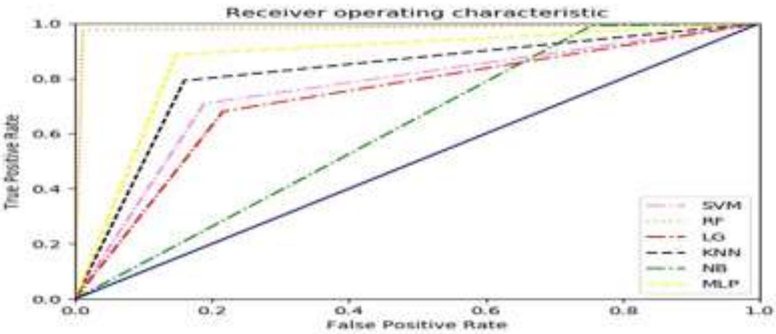

Fig. 2. ROC curve for the dataset with 9057 tweets

1) Dataset 1 with 9057 tweets: From this Table II , we can analyse that the algorithm **Random Forest** have the highest accuracy during majority of feature extraction methods except for the Count Vectorizer in which **MLP** attained the highest accuracy. And also when pragmatic and Syntactic feature was applied, almost all algorithms attained similar accuracy.

In the next step we concatenated all the features togetherto form the final feature array and checked whether accuracy has been improved or not.

An easy way to visualize the metrics is by creating a ROC curve, which is a plot that displays the sensitivity and specificity of a model. The true positive rate represents the proportion of observations that are predicted to be positive when indeed they are positive.Conversely, the false positive rate represents the proportion of observations that are predictedto be positive when they're actually negative.

From the Fig. 2, we can understand that Model **Random Forest** has the highest AUC, which indicates that it has the highest area under the curve and is the best model at correctly classifying observations into categories followed by KNN and MLP.And also we can come to the analysis that **Naive Bayes** attained the lowest accuracy in all of the algorithms.

2) Dataset 2 with 22890 tweets: As already mentioned, our dataset had 20001 tweets initially with 12179 non-bullying (positive) tweets and 7822 bullying (negative) tweets. Thedataset was imbalanced with non bullying class as majority. In order to rectify that, smaller datasets and tweets available from public sources were added , thereby reducing the imbalance and resulting in the dataset of 22890 tweets.

The Fig. 3 shows the accuracy of the imbalanced dataset or the raw twitter dataset we got from public source.We can seenthat RF outperforms the other algorithms, followed by KNN and MLP with only slight differences.

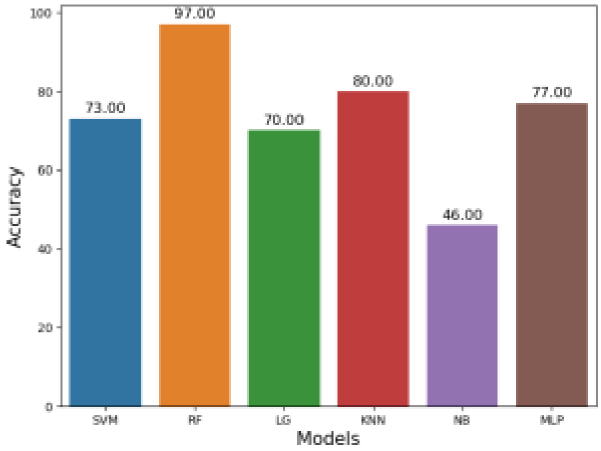

Fig. 3. Accuracy for the imbalanced dataset of 20001 tweets

By analysing the Fig.4, we can understand that accuracyhas been improved for the balanced dataset in most of the algorithms. And Random forest gives the best accuracy.

The more that the ROC curve hugs the top left corner of the plot, the better the model does at classifying the data into categories.To quantify this, we can calculate the AUC (area under the curve) which tells us how much of the plot is locatedunder the curve.The closer AUC is to 1, the better the model.Here also from the Fig. 5, the conclusion can be made that Model **Random Forest** has the highest AUC on both datasets, thus being the best model at correctly classifying observations into categories. The only variation found here is that the second best model is MLP followed by KNN.

An analysis was also done on the basis of impact of each feature to the final accuracy. This was attained by checking the accuracy achieved on the removal of each feature.This analysis gave the picture of which feature contributed the mostto accuracy or which can be considered as the base feature.

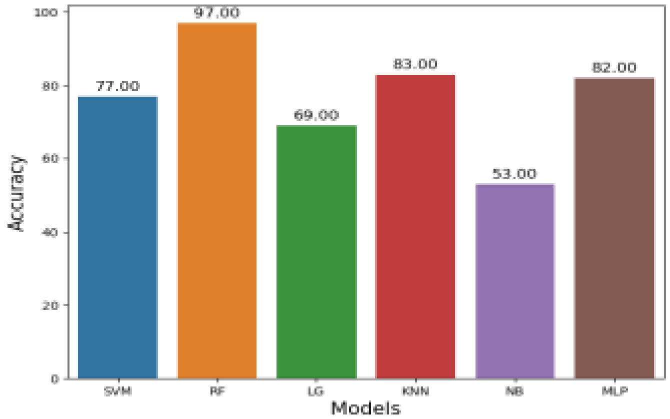

Fig. 4. Accuracy for the dataset of 22890 tweets

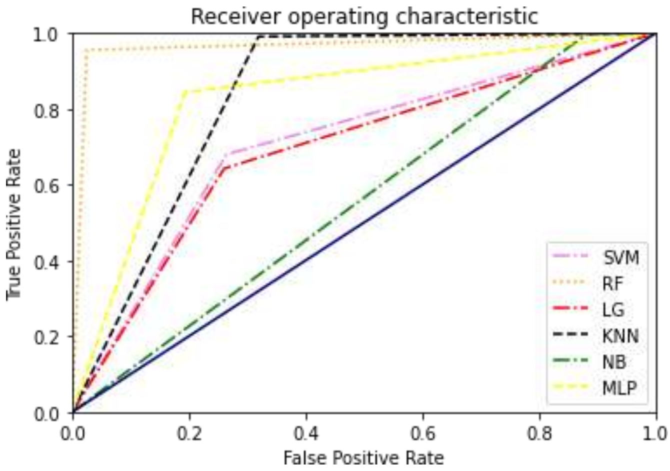

Fig. 5. ROC curve for the 22890 dataset

TABLE III
ACCURACY ACHIEVED AFTER REMOVING EACH FEATURE

Models *A*	**Original Accu-racy** *B*	**TFIDF** *C*	**Sentiment** *D*	**Semantic** *E*	**Pragmatic Syntactic** *F*	**-Count Vector-izer** *G*
SV M	**77**	76	*72*	74	75	75
RF	**98**	98	*89*	93	97	98
LG	**74**	74	*72*	73	73	74
KNN	**83**	82	*78*	77	82	83
NB	**57**	59	*61*	60	58	59
MLP	**84**	85	*82*	81	85	84

From the table III,we can understand what difference it can make to the overall accuracy of the system if we remove each of the features one by one. Only when the Sentiment and Semantic features are removed, the overall accuracy drops drastically. TFIDF and Count Vectorizer has very small impact on the final accuracy, So even if its removed, it does not causemuch change.

C. Performance Metrics

Performance measures generally evaluate specific aspects of the performance of classification tasks and do not always present the same information. Understanding how a model performs is an essential part of any classification algorithm. The underlying mechanics of different evaluation metrics may vary, and for comparability it is crucial to understand what exactly each of these metrics represents and what typeof information they are trying to convey. There are several methods to measure performance of a classifier: example met- rics are recall, precision, accuracy, F-measure, micro-macro averaged, precision and recall [6]. These metrics are based on "Confusion Matrix" that includes

- true positive (TP): the number of instances correctly labelled as belonging to the positive class
- true negative (TN): negative instances correctly classifiedas negative
- false positive (FP): instances incorrectly labelled as be-longing to the class
- false negative (FN): instances that are not labelled asbelonging to the positive class but should have been.

Precision, **Recall** and **F measure** are the metrics we have used to evaluate machine learning algorithms since accuracy alone is not sufficient to understand the performance of classificationmodels.

1) Precision

- This quantifies the number of positive class pre- dictions that actually belong to the positive class. Precision, therefore, calculates the accuracy for the minority class.
- Higher precision means that an algorithm returns more relevant results than irrelevant ones and it relates to the low false positive rate.
- The precision is the ratio of

$$\frac{tp}{(tp + fp)}$$

where tp is the number of true positives and fp isthe number of false positives

- From the above Fig 6,we can understand that high precision are obtained in almost all the models in both the datasets.
- We have got 0.69 precision as the lowest value for Logistic Regression in the dataset of 22890 tweets, which we consider as "good" and highest precision value of 0.98 and 0.97 is attained by Random Forest in 9057,22890 datasets respectively.

2) Recall

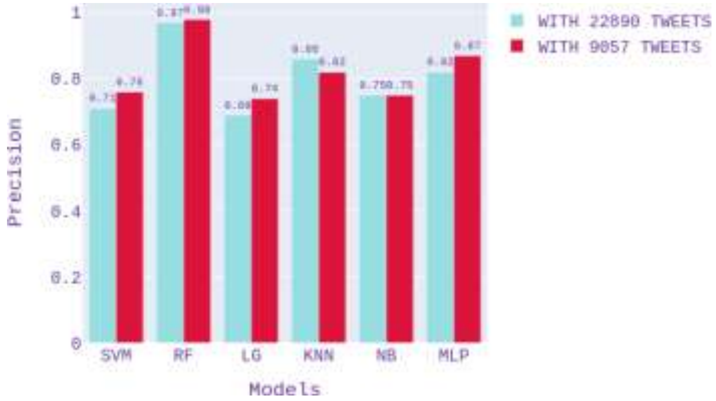

Fig. 6. Precision with both datasets

- At the same time, recall of Naive Bayes is very low which means most the predictions of Naive Bayes are correct, but most ground-truth objects have not been detected (many false negatives).

3) F-Measure

- This provides a single score that balances both the concerns of precision and recall in one number.
- A good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real bullying tweets and you are notdisturbed by false tweets.
- F1 Score is calculated as

$\underline{(2 \ast \text{Precision} \ast \text{Recall})}(\text{Precision} + \text{Recall})$

TABLE IV F1-SCORE FOR THE TWO DATASETS

Models	Dataset of 9057 tweets	Dataset of 22890 tweets
SV M	76	71
RF	98	97
LG	73	69
KNN	82	82
NB	54	44
MLP	87	82

- This quantifies the number of positive class pre- dictions made out of all positive examples in the dataset.
- Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions.
- The recall is the ratio of

$$\frac{tp}{(tp + fn)}$$

where tp is the number of true positives and fn thenumber of false negatives

- From the table IV, we can understand that **Random Forest** has the highest F1-Score followed by MLP and KNN for both the datasets.
- The lowest F1-Score is for the Naive Bayes , thereby we can conclude its not an ideal detection model forcyberbullying.

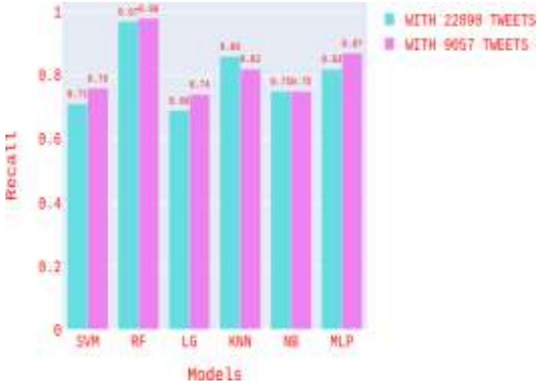

Fig. 7. Recall with both datasets

- From the figure 7, we can understand that the recall value of Random Forest is higher which means it implies an ideal detector that has detected majority results correctly.

VIII. CONCLUSIONS

This project is an approach to detect cyberbullying from Twitter social media platform based on different feature analysis like TFIDF, CountVectorizer, Sentiment, Semantic ,Syntactic and Pragmatic features that employed six ma- chine learning techniques; namely, Random Forest,Naive Bayes, Support Vector Machine,Logistic Regression, K-Nearest Neighbour,Multi-Layer Perceptron. Usage of these additional features have increased the overall accuracy of the model. The two Twitter datasets with 9057 and 22890 tweets used was collected from Git repository and is a collection of tweets that have been classified into positive (bullying) and negative(non-bullying).Before training and testing with machine learning models, the collected set of tweets have gonethrough several phases of preprocessing steps like cleaning, tokenization, stop words removal and lemmatization etc.

The results of the conducted experiments have indicated that Random Forest have outperformed all the other classifiers in all performance measures over all the two datasets we tried. And also Naive Bayes has obtained the lowest accuracy of all. The performance was measured with different metrics like precision, recall and F1-Score.All the experiments showed the same result as Random Forest as the best model for detection.

In all the studies we checked in literature survey, it was mentioned that SVM outperforms all the other algorithms, but we got a different scenario as the result.

Finally, for direction research in cyberbullying detection, we would like to explore other machine learning techniques such as Neural Networks and deep learning, with larger sets of tweets and also more feature extraction methods including user features, network features and social features.

IX. LIMITATIONS AND FUTURE SCOPE

A. Limitations

We could not perform in depth analysis in relation to users' behavior because the dataset we used for this study did not pro-vide any information (i.e. time of the tweet, favorite, followersetc.) other than just content (tweets). Moreover, we could have performed the meta-analysis on the effects of cyberbullying severity, however, also because the studies that we reviewed did not provide necessary information that would enable this type of analysis. Despite these limitations, we believe that the present work can be considered as a stepping stone for anybody who works for identifying cyberbullying severity into different levels to build machine learning multi-classifier. Furthermore, present study is only focused on twitter. Other social network platforms (such as Facebook, YouTube etc) alsoneed to be investigated to see the same pattern of cyberbullyingseverity.

B. Future Scope

For future direction research in cyberbullying detection, we would like to explore other machine learning techniques such as Neural Networks and deep learning, with larger sets of tweets and also with more feature extraction methods including user features, network features and social features. And also if we could implement ensemble modelling, it would be better to improve accuracy. The ensemble model can aggregate the prediction of each base model (SVM,Random Forest,Logistic Regression,Multi-Layer Per- ceptron,Naive Bayes,KNN) that we have and results in one final prediction.Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets.

REFERENCES

[1] M. D. Capua, E. D. Nardo, and A. Petrosino. Unsupervised Cyber Bullying Detection in Social Networks. 23rd International Conference on Pattern Recognition (ICPR), 2016.

[2] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringh- ini, and A. Vakali. Mean birds: Detecting aggression and bullying on twitter. In Proceedings of the 2017 ACM on web science conference, pages 13–22, 2017.

[3] Y. Chen, S. Zhu, Y. Zhou, and H. Xu. Detecting offensive language in social media to protect adolescent online safety, 2012.

[4] M. Dadvar and F. de Jong. Cyberbullying detection: a step toward a safer internet yard. In Proceedings of the 21st International Conference on World Wide Web (WWW ’12 Companion)ACM Digital Library. and Hypermedia and Web. Association for Computing Machinery. Special Interest Group on Hypertext, 2012.

[5] R. K, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying, 2011.

[6] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, andD. Brown. Text classification algorithms: A survey, 2019.

[7] G. A. León-Paredes, W. F. Palomeque-León, P. L. Gallegos-Segovia, P. E. Vintimilla-Tapia, J. F. Bravo-Torres, L. I. Barbosa-Santillán, and M. M. Paredes-Pinos. Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the spanish language. In 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), pages 1–7, 2019.

[8] P. M and D. J. Cyberbullying: Experiences, impacts and coping strategies as described by australian young people. Youth Studies Australia., 2010.

[9] Nandakumar, V. Kovoor, B. C, and S. M. U. Cyberbulling revelation in twitter data using naive bayes classifier algorithm. International Journal of Advanced Research in Computer Science, 9, 1 2018.

[10] J. W. Patchin and S. Hinduja. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. Youth Violence and Juvenile Justice, 4:148–169, 2006.

[11] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados. Detecting and monitoring hate speech in twitter. Sensors (Switzerland), 19, 11 2019.

[12] V. C. o. E. Rahul Ramesh Dalvi;Sudhanshu Baliram Chavan; Aparna Halbe, I. of Electrical, and E. Engineers. Detecting A Twitter Cyberbullying Using Machine Learning. Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) : 13-15 May, 2020.

[13] R. I. Rasel, N. Sultana, S. Akhter, and P. Meesad. Detection of cyber- aggressive comments on social media networks: A machine learning and text mining approach. In Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval, pages 37–41, 2018.

[14] S. Salawu, Y. He, and J. Lumsden. Approaches to automated detection of cyberbullying: A survey. IEEE Transactions on Affective Computing,11:3–24, 1 2020.

[15] A. Singh and M. Kaur. Detection framework for content-based cyber- crime in online social networks using metaheuristic approach. Arabian Journal for Science and Engineering, 45, 9 2020.

[16] R. E. Trana, C. E. Gomez, and R. F. Adler. Fighting cyberbullying: An analysis of algorithms used to detect harassing text found on youtube. In T. Ahram, editor, Advances in Artificial Intelligence, Software and Systems Engineering, pages 9–15, Cham, 2021. Springer International Publishing.

[17] N. Tsapatsoulis and V. Anastasopoulou. Cyberbullies in twitter: A focused review. In 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pages 1–6. IEEE, 2019.

[18] J. Yadav, D. Kumar, and D. Chauhan. Cyberbullying detection using pre- trained bert model. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pages 1096–1100, 2020.