Using Random Forest and Naïve Bayes Algorithms in Detection of Cyberbullying on Twitter

Salisu Suleiman
Dept. of Computer Science Education
Isa Kaita College Education Dutsin-ma
Katsina State of Nigeria
Email: salisusumar@gmail.com
&

Tanimu Lawal
Dept. of Management & info. Science
Yusuf Bala Usman College Daura
Katsina State of Nigeria
Email: ltanimulawal@gmail.com

Abstract
The internet has infiltrated every aspect of human life, making it simpler to connect people all over the world and share information to a wider range of people. The purpose of this research is to carry out a comparative analysis and performance evaluation of both machine learning algorithms used in this study for the detection of bullying tweets.Nowadays, cyberworld has numerous negative impacts on individuals despite its significant importance. One of the most dangerous threats in the cyberworld is cyberbullying as it destroys individuals' reputation or privacy, threatens or harasses them, and sometimes leads to suicidal acts. Therefore, an effective and automatic detection model is proposed so that the bullies' abusive tweets or insulting comments can be identified and detected using machine learning and natural language processing. Two machine learning algorithms in this research, viz: Naïve Bayes (NV) and Random Forest (RF) were used for the cyberbullying detection. The datasets retrieved from Kaagle was used to train and test the model for classification of tweetSets as either bullying or non-bullying (binary class classification model) and the features were extracted using Term Frequency-Inverse Document Frequency (Tf-idf).

Keywords: Cyberbullying, Cybersecurity, tweetSets, Machine Learning (ML), Natural Language Processing (NLP).

Introduction

Cyberbullying is an aggressive, premeditated action conducted through the Internet against a defenseless individual or other electronic means like emails, text messages, social media messaging, or web site/blog content (Ozel et al., 2017). Cyberbullying could also be described as tyranny carried out online via electronic devices such as computers, mobile phones, and tablets. It can occur through text messages or on the internet in forums, social media platforms, or video games where members of the community can share and express their opinions. In short, bullies used social media to harass people. Bullying can be understood as a

set of behaviour with the intent to injure, which leads to suicidal thoughts, emotional reactions, and low self-esteem in the victim like frustration, fear, anger and depression (Islam et al., 2020).

At a teenage age, children in today's society want their own mobile phones and tablets, and they want to connect to social media platforms and play online games like Fortnite. If their behaviour is left unchecked by their parents, it may result in cyberbullying. Rumors spread on social media or over e-mail; humiliating films or images; and insulting, frightening, and abusive words spread on social media are all examples of cyberbullying. When a message or a picture is shared on the internet, it is very difficult to track and remove the content from the social media. This may happen 24 hours a day, seven days a week and it can reach out to its victim when they are alone and away from home. Cyberbullying differs from traditional bullying in that the perpetrator does not have to physically confront his/ her victims. Some of the prominent definitions of cyberbullying are: "An aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself (Smith et al., 2008).

Cyberbullying occurs when someone continually makes fun of another person online, targets another person through e-mail or text messaging, or writes something negative about another person online" (Hinduja & Patchin, 2012). Sending harsh messages or DMs to someone, pranking people's calls, harassing someone in an online game, hacking into someone's social networking profile or game, spreading rumours about individuals online, and posing as someone else to distribute hurtful words online are all examples of cyberbullying. The victims can have a variety of negative consequences, including being rejected by their peers (leading to loneliness and isolation), low self-esteem, despair, poor academic performance, and decreased emotional well-being (Cowie, 2013). Individuals who have been cyberbullied are also more prone to have headaches, stomach pain, and insomnia (Sourander et al., 2010). In some extreme cases such as the Megan Meier incident, a 12-year-old girl who committed suicide after being bullied on the social media platform.

Moreover, Cyberbullying affects the victim in terms of emotional, psychological, and physical distress. Emotional effects involve isolation/anger (16%). Mental effects involve depression (22%), social anxiety (8%), self-harm (10%), suicidal thoughts (12%). Physical effects involve headache/sleeping disturbances (18%), eating disorders (14%) (Singh & Sharma, 2021). The impact of cyberbullying on an individual in an online social media network is depicted in the Figure 1:

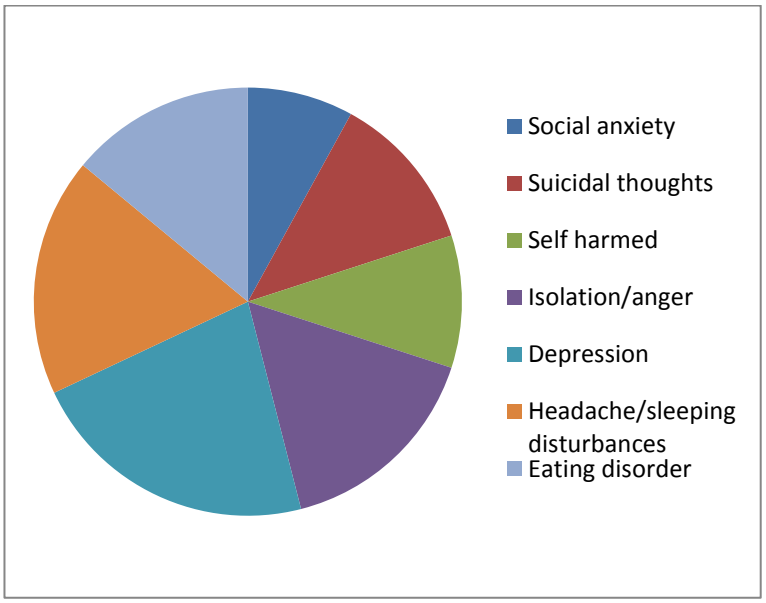


Figure 1: Cyberbullying impact on the victims

Social media can be termed as a platform that permits users to share anything they desire, like; images, documents and videos as well as communicate with (Islam et al., 2020). People use social media platforms to interact with one another via computers or cellphones.

Facebook, Twitter, Instagram, TikTok, and other social media platforms are among the most widely used worldwide. Currently, social media is used in a variety of fields, including education (Selwyn & Stirling, 2015), business and philanthropy. Social media is also boosting the world's economy through creating many new job opportunities (Dixit & Kumar, 2015). Victims of cyberbullying commit self-destructive actions such as suicide. In this era of web 4.0, where people live on digital and online platforms, it is highly difficult to protect society from the frightening rise of cybercrime. Consequently, recognizing and preventing cyberbullying is crucial in order to protect teenagers.

1. Literature Review

Cyberbullying has become a significant concern on social media platforms like Twitter. The anonymous nature of online interactions can facilitate bullying behavior, which can have severe consequences for the victims. Machine learning algorithms, such as Random Forest and Naïve Bayes, have been employed to detect cyberbullying on Twitter.

A Random Forest-based approach was developed by (Novalita et al., 2019) for detecting cyberbullying on Twitter, achieving an accuracy of (93.6%).

A comparative analysis was conducted by (Chakraborty *et al.*, 2020) on the performance of Random Forest, Naïve Bayes, and Support Vector Machine (SVM) algorithms for

cyberbullying detection on Twitter. Random Forest outperformed the other algorithms, achieving an accuracy of (92.5%).

Random Forest algorithm was used by (Arora et al., 2019) for cyberbullying detection on Twitter, reporting a precision of (91.4%) and recall of (90.5%).

In order to detect cyberbullying on Twitter, (Nandakumar, 2018) proposed a Naïve Bayes-based technique that achieved an accuracy of (88.2%).

With a precision of (86.5%) and recall of (85.2%), (Kumar A. et al., 2020) used Naïve Bayes to detect cyberbullying on Twitter.

The performance of the Random Forest and Naïve Bayes algorithms for Twitter cyberbullying detection was compared by (Roy, P. et al., 2022) Naïve Bayes was surpassed by Random Forest, which achieved (92.1%) accuracy. Random Forest outperformed Naïve Bayes with an accuracy of (91.8%) in a comparison of the algorithms' performance for detecting cyberbullying on Twitter by (Kumar . *et al.*, 2020).

According (Islam et al., 2020), four machine learning algorithms such as: Support Vector Machine, Random Forest, Naïve Bayes and Decision Tree were used to identify bullying content on social media in English using two features i.e; Bag of Words (Bag of Words) and Term Frequency-Inverse Document Frequency (TF-IDF) to analyse the level of accuracy of four Machine learning algorithms used. Facebook and twitter dataset were successfully downloaded from kaggle.com. The result indicated that SVM outperforms all other machine learning classifiers used in the research. In the same way, Term Frequency-Inverse Document Frequency (TF-IDF) outperformed Bag of Words (BoW).

The majority of cyberbullying detection researches according to (Pawar, R. et al., 2019) were done in a single language. As a result of the damaging nature of cyberbullying, a model capable of identifying cyberbullying in multiple languages, such as Hindi and Marathi is developed. The datasets were collected from different sources, including newspaper reviews, tourist reviews collected manually, and tweets retrieved from the Twitter API. The results reveal that the F1-scoreachieved (97%), and accuracy was measured to be (96%). In both Hindi and Marathi, the percentages were calculated. In all three datasets examined, Logistics Regression (LR) beats Stochastic Gradient Descent (SGD) and Multinomial Naïve Bayes (MNB).

In a study conducted by (Jain, V. et al., 2021) four machine learning algorithms were used to detect bullying text in English, including Support Vector Machine (SVM), Logistics Regression (LR), Random Forest (RF), and Multilayered perceptron algorithms, as well as three distinct textual features such as BoW, Word2Vec, and Tf-idf, and the dataset was obtained from Wikipedia and Twitter. The findings indicated that the Twitter dataset had more than (90%) accuracy and the Wikipedia dataset had more than (80%) accuracy when applying the same machine learning classifiers. The BoW and Tf-Idf features considerably outperformed the Word2Vec function.

Support Vector Machine and Naïve Bayes are two machine learning algorithms that are used for detecting bullying tweets on Twitter and Wikipedia in both Arabic and English. (Haidar, B. et al., 2017 October) indicated that it is important to detect cyberbullying on different social media platforms. NV outperformed SVM on both the Twitter and Wikipedia datasets with an accuracy of (90.8%). Random Forest (RF), AdaBoost (ADB), Naïve Bayes (NB), Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM) and as machine learning classifiers were employed by (Muneer, A., & Fati, S. M., 2020) Each of these methods was tested using accuracy, precision, recall, and f1-score as performance metrics to determine the classifiers' recognition rates applied to the global dataset. A global dataset of 37,373 was used to test the seven classifiers used in the study. Among the classifiers, logistic regression got the greatest f1-score value of (0.928), Stochastic Gradient Descent (SGD) had the best precision of (0.968), and Support vector Machine (SVM) had the best recall of (1.00). Finally, with a median accuracy of around (90.57%), the studies revealed that Logistic Regression is superior.

In a study conducted by (Zhang et al. 2020), Linear Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, Perceptron, and Logistic Regression techniques were used to identify cyberbullying in Japanese texts. The dataset was obtained from Twitter and shows a fairly even distribution. The Gradient Boosting approach produces the greatest results, with an f1-Score up to (93.5%).

Methodology

In this section, the proposed method for cyberbullying detection is explained. The datasets used is also presented; the performance metrics as well as the algorithms used are also described. Collection of raw datasets, Natural Language Processing (NLP), Machine Learning Model, and Result Analysis are all part of the proposed framework for detecting cyberbullying depicted in figure 2:

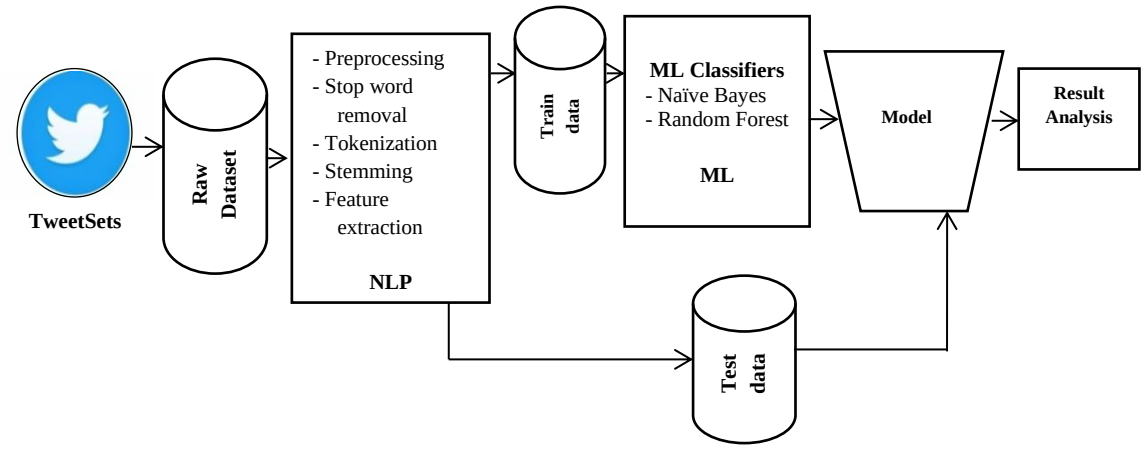


Figure 2: Proposed frame work for cyberbullying detection

i. Dataset

In machine learning, a dataset is essentially a collection of data bits that can be analyzed and predicted by a computer as a single entity. This means that the data collected should be made uniform and understandable for a machine that doesn't see data the same way as humans do.

In this research, we use Dataturks‟ Tweet Dataset for Cybertroll Detection which would be obtained from Kaggle (DataTurks, 2018) for reaching the final results. Because of the seriousness of the issue we aim to resolve, it is crucial to choose a dataset that is complete, reliable, relevant, and to the point.

ii. Data Collection

As previously stated, the Dataturks‟ Tweet Dataset for Cybertroll Detection would be collected from Kaggle for our model's training and testing. We initially took into account a large number of additional datasets, but upon thorough examination, it was discovered that many of them had irrelevant data, had missing attributes, or were of poor quality. Therefore, after experimenting with several different publicly available datasets, we chose (DataTurks, 2018) as it appeared to meet all of our requirements.

iii. Data Cleaning

Datasets usually come with irrelevant characters, noise data or missing attributes, so, once the datasets are collected, we prepare and clean them for training and testing of our model.

iv. Data Preprocessing

The preprocessing steps are done as follows:

1) Word Tokenization: A token is a single entity that serves as a phrase or paragraph's building block. Word Tokenization breaks down our text into individual words in a list.

2) Stop Words Removal: This is to fetch a list of stopwords in the English dictionary, after which they are removed. Stop words are non- significant terms like "the," "a," "an," and "in" that have no bearing on the interpretation of the data to be evaluated.

3) Punctuation removal: Only characters that are not punctuation are saved here, which are verified using string.punctuation.

4) Stemming: A linguistic normalisation procedure in which terms are reduced to their underlying word. Tokens are stemmed using nltk.stem.porter.PorterStemmer to get the stemmed tokens. For example, the following words; "connection," "connected," and "connecting" may all be

reduced to aroot word "connect."

5) Digit removal: Any numerical data are removed since it does not contribute to cyberbullying.

6) Feature Extraction: The next step is to extract features so that they could be utilised with machine learning techniques, which would be done using Python's sklearn library and the Term Frequency-Inverse Document Frequency (TF-IDF) Transformer. The TF-IDF is a statistical method for determining the significance of a word. The number of times the word appears in the document is multiplied by the inverse of the term's document frequency. Rather than calculating the frequency of words like CountVectorizer does, TF-IDF uses a technique that reduces the weight (importance) of words that exist in many texts in common, deeming them incapable of distinguishing the documents.

v. Machine Learning

Machine Learning (ML) is defined as the ability of a computer to teach itself how to take a decision using available data and experiences (Muneer, A., & Fati, S. M., 2020). The Data is known as *Training Data*. Decisions to be taken in ML might be classification or prediction. The computer classifies a new piece of data by training models using learning algorithms. In order to detect cyberbullying from social media texts, some typical machine learningclassifiers employed in the research are discussed below:

A. Gaussian Naïve Bayes

Gaussian Naive Bayes classifier is a collection of classification algorithms based on Bayes‟ Theorem of mathematics. The Bayes' theorem, in basic words, determines the likelihood of an event occurring based on prior knowledge of factors that may be important to the event. It's a group of algorithms that all work on the same premise: that each pair of classified features is independent of the others. For binary (two-class) and multi-class classification issues, Naïve Bayes is an appropriate classification algorithm. The technique is easier to grasp when expressed in binary or categorical input values. It is so named because the calculation of the probability for each hypothesis is reduced to make it tractable. By assuming a Gaussian distribution, Gaussian Nave Bayes is usually extended to real-valued features. This extension of Naïve Bayes is called Gaussian Naïve Bayes. There are also Multinomial Naïve Bayes and Bernoulli Gaussian Naïve Bayes, in addition to Gaussian Naïve Bayes. We picked Naïve Bayes because it is the most widely used one and one of the simplest to implement because we only need to estimate the mean and the standard deviation from the training data. The classifier is implemented using sklearn.naiveBayes_package.

B. Random Forest Classifier

As the name suggests, the Random Forest Classifier is composed of a vast number of separate decision trees that collaborate as a group. Every tree in the random forest generates a class prediction, and our model predicts the class with the most votes. Because the trees shield one another from their individual flaws, the models' low correlation enables them to provide ensemble forecasts

that are more accurate than any single prediction. The classifier is implemented using sklearn.ensemble package.

2. Experimental Results

In this experiment, the two machine learning algorithms were run on the Dataturks‟ Tweet Dataset for Cybertroll Detection obtained from Kaggle for detection of cyberbullying content. Random Forest classifier outperformed Naïve Bayes classifier with an accuracy of (92%) and F1-score of (0.92) whereas Naïve Bayes was the poorest with an accuracy of (62%) and average of (0.62).

Accuracy, Precision, Recall and F1-score were used in our experiment as the performance metrics to determine the performance of every classifier. These metrics utilize True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Their computational formulae are indicated below - "T" in these formulae indicated the summation of TP, FP, TN, and FN.

The performance metrics are briefly explained as follows:

1. Accuracy: Accuracy measures the amount of accurate or correct predictions made by themodel. It is formulated as:

   Accuracy = (TP + TN) / T

2. Precision: Precision is the measure of bullying tweets correctly predicted by the algorithm. It is formulated as:

   Precision = TP / (TP + FP).

3. Recall: Recall is the ratio of how many bullying tweets, out of all available ones, are actually detected by the algorithm. It is formulated as:

   Recall = TP / (TP+FN).

4. F1-Score: F1-score gives an unbiased class-wise result. It takes both false positives and false negatives into account and returns the weighted average of Precision and Recall. It is calculated as:

   F1 = 2*((Precision* Recall)/(Precision + Recall)).

I. Result of Naïve Bayes algorithm

After running the Naïve Bayes algorithm, an accuracy of (62%) and average score of (0.62) were achieved. The detail of the result is shown in table 1:

Table 1: Result of Naïve Bayes Classifier

Annotation	Precision	Recall	F1-score
0	0.97	0.38	0.55
1	0.51	0.98	0.67
Accuracy			0.62
Macro avg.	0.74	0.68	0.61
Weighted avg.	0.79	0.62	0.59

II. **Result of Random Forest algorithm**

Random Forest algorithm was also run and achieved a highest performance with an accuracy of (92%) and average score of (0.92). Table 2 below carries the detail of the result:

Table 2: Result of Random Forest Classifier

Annotation	Precision	Recall	F1-score
0	0.97	0.90	0.93
1	0.86	0.95	0.90
Accuracy			0.92
Macro avg.	0.91	0.92	0.92
Weighted avg.	0.92	0.92	0.92

Figure 3 depicts the graphical comparisons on the accuracy level achieved by every machine learning classifier used. While fig. 4 shows the achievement of each classifier interms of every performance metric.

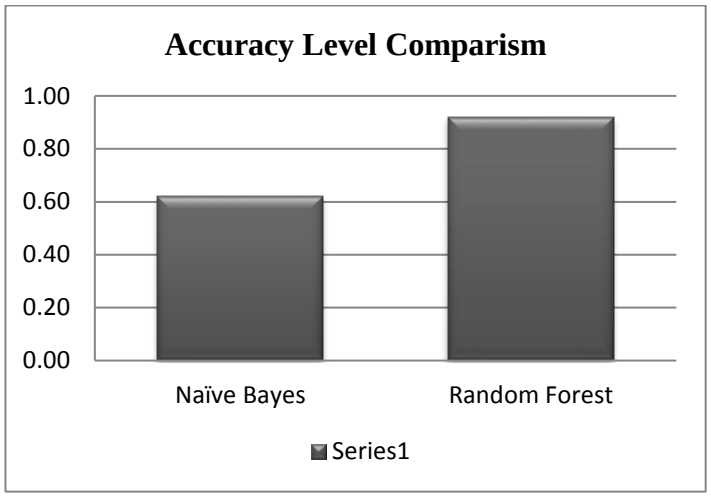


Figure 3: Graphical Representation of the Classifiers' Accuracy

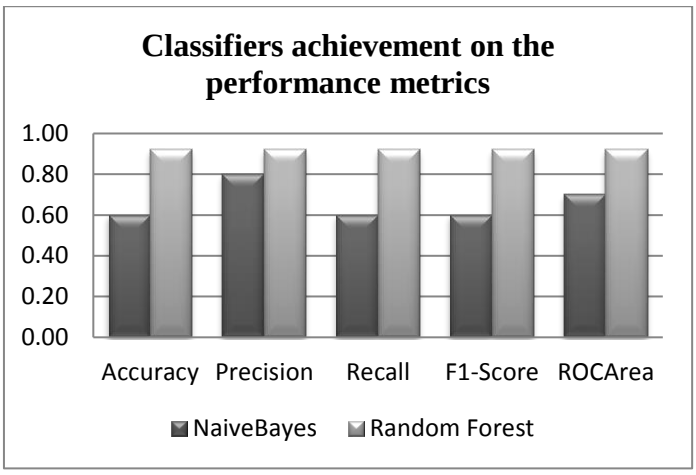
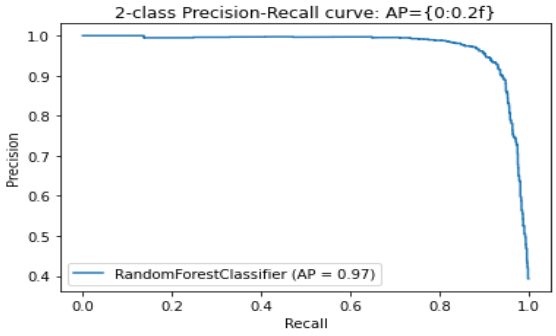


Figure 4: Precision, Recall, F1-Score & ROC Area



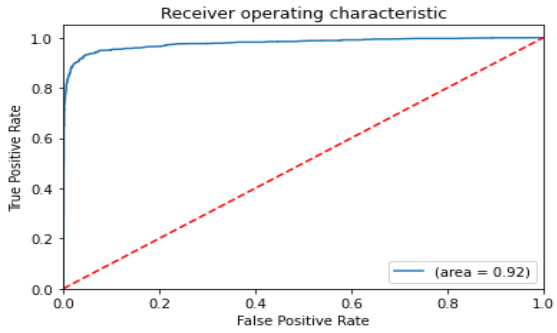


Figure 3: Naïve Bayes Classifier

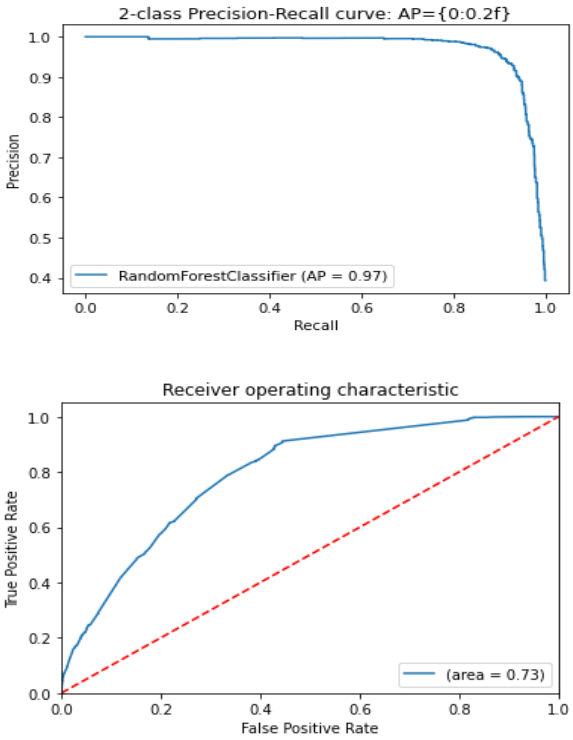


Figure 4: Random Forest Classifier

5. Conclusion and future work

Random Forest and Naïve Bayes algorithms have shown promising results in detecting cyberbullying on Twitter. Random Forest tends to perform better in terms of accuracy, precision, recall and F1-score with an overall performance of (92%) accuracy. Naïve Bayes performed the worst, giving just an approximated value of (62%) accuracy. However, both algorithms can be improved by combining them with other techniques and features.

For the future work, following are some observations made to improve the quality of the detection of cyberbullying content:

i. According to our extensive review on the related literatures, most of the researches done previously on cyberbullying detection, were text based, so our next target is the development of multimedia (image, audio and video) based detection model and this can be achieved by switching from conventional machine learning approaches to deep learningtechniques like Convolutional Neural Network CNN which are found good in dealing with any multimedia content.
ii. Our cyberbullying detection model is a binary class classification model (bullying or non- bullying), so multi-class classification approach could be also the direction of our future plan.

References

Ozel, S. A., Sarac, E., Akdemir, S., & Aksu, H. (2017). Detection of cyberbullying on social media messages in Turkish. *2017 International Conference on Computer Science and Engineering (UBMK)*, 366–370. https://doi.org/10.1109/ubmk.2017.8093411.

Islam, M. M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., & Acharjee, U. K. (2020). Cyberbullying detection on social networks using machine learning approaches. *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 1–6. https://doi.org/10.1109/csde50874.2020.9411601.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, *49*(4), 376–385. https://doi.org/10.1111/j.1469-7610.2007.01846.x.

Hinduja, S., & Patchin, J. W. (2012a). Cyberbullying: Neither an epidemic nor a rarity. *European Journal of Developmental Psychology*, *9*(5), 539–543. https://doi.org/10.1080/17405629.2012.706448.

Cowie, H. (2013). Cyberbullying and its impact on young people's emotional health and well-being. *The Psychiatrist*, *37*(5), 167–170. https://doi.org/10.1192/pb.bp.112.040840

Sourander, A., Brunstein Klomek, A., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., Ristkari, T., & Helenius, H. (2010). Psychosocial risk factors associated with cyberbullying among adolescents. *Archives of General Psychiatry*, *67*(7), 720. https://doi.org/10.1001/archgenpsychiatry.2010.79.

Singh, N., & Sharma, S. K. (2021). Review of Machine Learning Methods for identification of cyberbullying in social media. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 284–288. https://doi.org/10.1109/icais50930.2021.9395797.

Selwyn, N., & Stirling, E. (2015). Social Media and education … now the dust has settled. *Learning, Media and Technology*, *41*(1), 1–5. https://doi.org/10.1080/17439884.2015.1115769.

Dixit, M., & Kumar, H. (2015). *Tata McGraw Hill Education Private Limited*. https://doi.org/10.4135/9781473974463.

Novalita, N., Herdiani, A., Lukmana, I., & Puspandari, D. (2019). Cyberbullying identification on Twitter using random forest classifier. *Journal of Physics: Conference Series*, *1192*, 012029. https://doi.org/10.1088/1742-6596/1192/1/012029.

Chakraborty, T., & Ghosh, I. (2020). A Comparative Study of Machine Learning Algorithms for Cyberbullying Detection on Twitter. IEEE Transactions on Computational Social.

Arora, T., Sharma, M., & Khatri, S. K. (2019). Detection of cyber crime on social media using random forest algorithm. *2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC)*, 47–51. https://doi.org/10.1109/peeic47157.2019.8976474.

Nandakumar, V. (2018). Cyberbullying revelation in Twitter data using naïve Bayes classifier algorithm. *International Journal of Advanced Research in Computer Science*, *9*(1), 510–513. https://doi.org/10.26483/ijarcs.v9i1.5396.

Kumar, A., Yadav D. (2019). Detection of Cyberbullying in Twitter Using Naive Bayes Algorithm. Journal of Intelligent Information Systems, 54(2), 257-269.

Roy, P. K., Singh, A., Tripathy, A. K., & Das, T. K. (2022). Cyberbullying detection: an ensemble learning approach. *International Journal of Computational Science and Engineering*, *25*(3), 315-324.

Pawar, R., & Raje, R. R. (2019). Multilingual cyberbullying detection system. In *2019 IEEE international conference on electro information technology (EIT)* (pp. 040-044). IEEE.https://doi.org/10.1109/EIT.2019.8833846.

Jain, V., Kumar, V., Pal, V., & Vishwakarma, D. K. (2021, April). Detection of Cyberbullying on Social Media Using Machine learning. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1091-1096). IEEE. DOI: 10.1109/ICCMC51019.2021.9418254.

Haidar, B., Chamoun, M., & Serhrouchni, A. (2017, October). Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content. In *2017 1st cyber security in networking conference (CSNet)* (pp. 1-8). IEEE.https://doi.org/10.1109/CSNET.2017.8242005.

Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, *12*(11), 187. DOI:10.3390/fi12110187.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2020). Deep Learning based RecommenderSystem. *ACM Computing Surveys*, *52*(1), 1–38. https://doi.org/10.1145/3285029.

DataTurks. (2018, July 12). *Tweets dataset for detection of cyber-trolls*. Kaggle. https://www.kaggle.com/datasets/dataturks/dataset-for-detection-of-cybertrolls.