International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume 10 Issue 02, Feb 2023 www.irjet.net p-ISSN: 2395-0072

Depression Detection Using Various Machine Learning Classifiers

Amaan Khan, Jayeshkumar Kharol, Abhishek Gaikwad, Swapnil Jagtap, Prof. Suresh Mestry

Rajiv Gandhi Institute of Technology, Mumbai 400053
Professor Suresh Mestry, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Maharashtra, India

---***---

Abstract - Datasets derived from social networks are useful in several areas, including neuroscience and psychology. Recent studies have indicated a correlation between high usage of social media and increased depressed users. However, technical assistance is inadequate, and precise approaches are desperately needed. By using Machine learning and Natural language processing, depressed consumers of Twitter are detected. Sentiment analysis is used for obtaining the polarity of the text. The model is trained and tested using different forms of classifiers. The quality of each classifier is investigated and the pre-eminent model is used. Finally, within the proposed model, a GUI for depression detection is created. Hence the project would identify users with depression.

***Key Words*: Sentiment Analysis, depression detection, Machine learning (ML), Social Media , Natural Language Processing**

1. INTRODUCTION

According to the World Health Organization (WHO), depression is the leading cause of disability worldwide, affecting more than 264 million people of all ages worldwide. Post-traumatic stress disorder (PTSD) is a form of depression that develops after exposure to a potentially traumatic event, such as witnessing an accident, physical or sexual assault, combat, or death. Approximately 8 million adults in the United States will experience PTSD in a given year. It is a serious medical condition causing persistent thoughts of traumatic memories, insomnia, and severe anxiety. Early diagnosis and treatment are therefore of the utmost importance. Therapies and medications are available for effective treatment. With the rise of social media, people tend to post most of their activities online as text and images, which largely reflects their mental health status. Detecting these indicators can help with early diagnosis and save lives in the future. Depression is one of the most common mental illnesses, but remains undiagnosed or untreated due to lack of awareness or denial. Severe depression can be prevented by early detection of symptoms and timely intervention and treatment.

Because users express their feelings through messages or comments on the Twitter platform, sometimes their tweets refer to emotional states such as "joy", "sadness", "fear", "anger" or "surprise". We collect this data to analyze various features from these tweets.

This study uses data collected from 20,000 tweets from different user profiles. A number of classification methods are used to determine the degree of depression, of which ETF (Extra Tree Classifier) shows the best results with an accuracy of 94 and precision of 97.29.The rest of the Article is as follows. Section 2 illustrates literature review .Section 3 explains the methodology of study and the features extracted. Section 4 represents the comparative study and system architecture. Section 5 describes the discusses result and conclusions . Finally Section 6 outlines the future work of the study

2. Literature Review

Psychological analysis can be carried out using various methods, with the text-based dataset. In this section, we discuss the previous works performed using various techniques for psychological analysis and the depression detection task. With the gradual increase in social media usage and the extensive level of self-disclosure within such platforms, efforts to detect depression from Twitter data have increased [1].

[2]Studies have been done on both imbalanced and balanced data. Any data that does not contain missing values in imbalanced data may be subject to resampling. Balanced data is divided into two types: (1) positive tweets and (2) negative tweets. In an imbalanced dataset, the number of instances with positive and negative tweets is checked first. This reduces the overall predictive performance of the model. Modelling of imbalanced data is done using data resampling methods.

Experiments in the proposed task were performed three times. Experiment 1 performs depression detection using an imbalanced data set. In Experiment 2, the SMOTE resampling method was implemented to solve the class imbalance problem of the training dataset. Experiment 3 implements the RUS under sampling method to solve the problem of imbalance in the data set. We analyze and compare the performance of the above classifiers. The LSTM classification model outperforms other basic models in its approach to detecting depression.

[3]The focus of the study was to combine four types of factors: affective processes, temporal processes, verbal styles, and all (emotional, temporal, and verbal styles) traits to detect and process depression data received in the form of Facebook posts. Study and detect depressive behavior in

Facebook comments. Information taken from social networks . NCapture was used to prepare social media data for Facebook[,] data collection. explored elements of each type individually using a supervised machine learning approach. For each type, classification methods such as "decision trees", "k-nearest neighbors", "support vector machines", and "ensembles" are considered appropriate. The results show that decision trees (DTs) outperform other machine learning approaches in a variety of experiments.

[4]A quantitative study was conducted to train and test different machine learning classifiers to determine if a user of a Twitter account is depressed based on tweets initiated or behavior on Twitter. Data preparation, feature extraction and classification tasks were performed in R version 3.3 with Rstudio IDE . To avoid overfitting, the classifier was trained using 10-fold cross-validation and then tested on an extended test set. They demonstrated depression detection using the Activity and Content Functional Classification (DDACF) model. First, all tweets for depressed and non-depressed accounts are retrieved, as well as user account and activity information such as number of followers, number of followers, total messages, post time, number of mentions, and number of likes.

[5]Quantitative studies were derived as review articles by comparing methodologies with those published in journals. Although these studies provide encouraging results, they also highlight some important limitations and challenges. For example, the accuracy of detecting depression in social media data is determined by the quality and representativeness of the data and the algorithms and features used in the analysis.

3. Methodology

People frequently experience mental health issues. Depression is the fastest-growing health disorder; it is caused by a change in mood, which includes elements of motivational and emotional conditions. Despite the popularity of social media platforms and the rapidity with which they have permeated almost every aspect of our lives, there is a significant lack of clear data on how they affect us personally, such as our behaviour, social relationships, and mental health. We conducted a quantitative study in this paper to train and test various machine learning classifiers to determine if a user of a Twitter account is depressed based on tweets initiated by the user or his/her Twitter activity.

3.1 Data preparation

Used data from 20,000 Twitter users. Data collected from social media platforms can contain errors or obsolete text, making sentiment analysis difficult. Since there are no emoticons in the data set we are using, there is no need to process emoticons. Feature extraction reduces the number of raw data processing groups.

Feature extraction is the process of selecting data and combining them into features to reduce the amount of data that must be accurately processed and carefully describe the original data set. The amount of redundant data for this analysis is also reduced. Therefore, feature extraction is performed by removing unnecessary data.Then clean the tweets by converting them to lowercase, removing punctuation and numeric values, and removing stopwords. Stopwords are important in many applications because they allow you to focus on the important words by removing commonly used words in a given language. Removing stopwords is done in Python using NLTK. A list of stopwords can be loaded, and the system is used to extract roots by removing suffixes or prefixes associated with words. In this study, a Snowball morpheme analyzer, which is different from the Porter morpheme analyzer, is used in that it can perform morpheme analysis on multiple languages. Tfidfvectorizer is used to tokenize the given document. Frequency Analysis is also carried out.

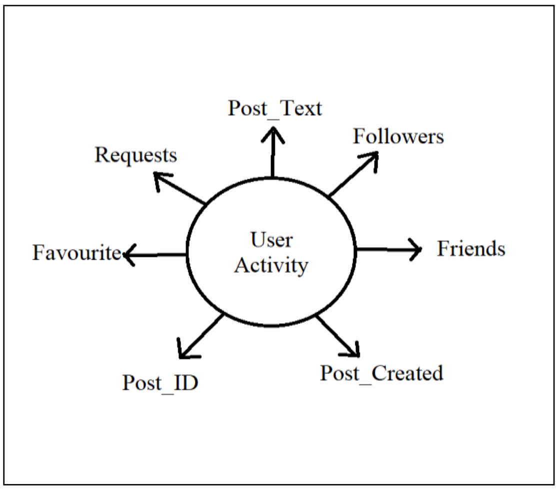


Fig 3.1 User Activity feature extracted from user account

3.2 Sentiment Analysis

Sentiment analysis (also known as opinion analysis) is a natural language processing (NLP) technique for determining whether data is positive, negative, or neutral. Sentiment analysis focuses on the polarity of the text (positive, negative, or neutral), but goes beyond polarity to specific feelings and emotions (anger, joy, sadness, etc.), urgency (urgent, non-urgent), and even intentions (interested vs. no interest). TextBlob was used to analyze the sentiment of the dataset.

TextBlob uses the Natural Language ToolKit (NLTK) and input is limited to one sentence. TextBlobs breed polarity and subjectivity. Polarity scores range (-1 to 1), with -1 identifying the most negative words such as "disgusting", "terrible" and "sorry" and 1 identifying the most negative words such as "excellent" and "best". Identify positive words. ". . The subjectivity score is between (0 and 1). It shows the degree of personal opinion. If the proposal is highly subjective, i.e. Closer to 1, the text appears to contain more personal opinions than factual information.

3.3 Modeling

The cleaned dataset is now used to train and test the machine learning model. Scikit-learn makes it easy to create and train models. The data set is split into two parts, training and testing, and is split 80/20. The dataset is trained with various classifiers such as Random Forest, Decision Tree, Naive Bayes, Regression, etc. and accuracy is calculated. We found that the Extra Tree had the highest accuracy in predicting whether or not a Twitter user was depressed.

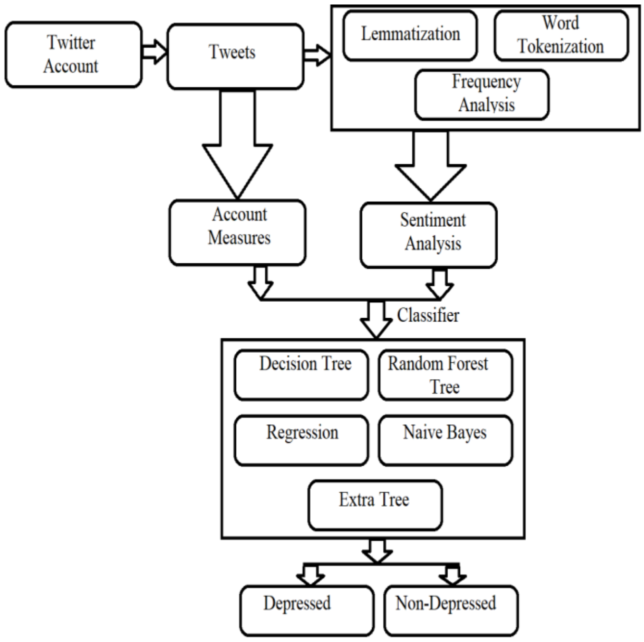


Fig 3.2 Classification Model

4 .0 Comparative Study & System Architecture:

The proposed action has been completed. We discuss, analyze and compare the work of classifiers. The classification algorithm was analyzed using the scikit-learn library and the plots were generated using the Matplotlib library.

The data was pre-processed and cleaned before identifying the most common words in the data set. As shown in the Fig 4.1

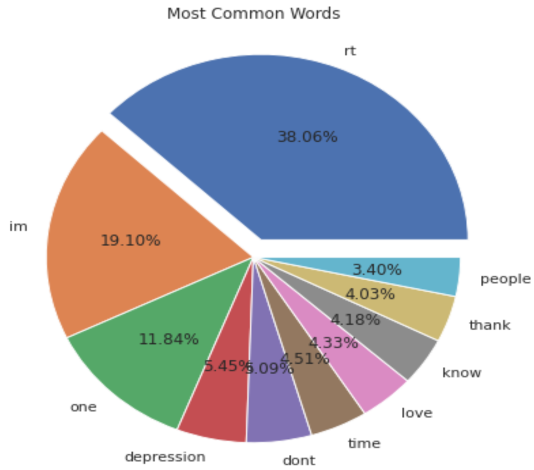


Fig 4.1 Pie chart of most common words

After applying Sentimental Analysis, we compared the positive tweets versus Negative tweets in the fig 4.2 below.

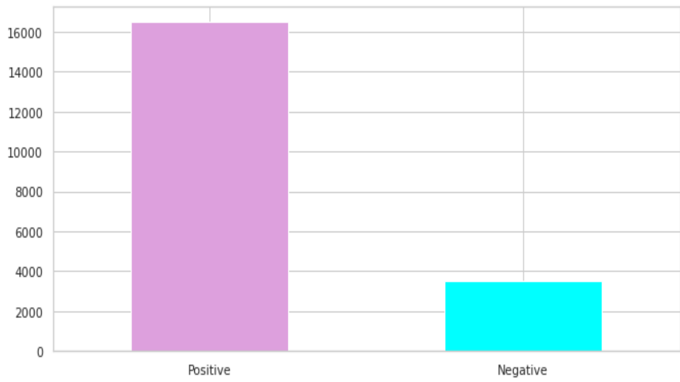


Fig 4.2 Sentiment Analysis Result

Word cloud was created to highlight popular words used in negative moods and phrases based on their frequency and relevance. This gave us quick and easy visual insights that could lead to deeper analysis.

Fig 4.3 Depressed Word Cloud

Different classifiers such as Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Decision Trees (DT), Complementary Tree Classifier (ETF), Random Forest (RF), Logistic Regression (LR), Nave Bayes, Gradient Boosted Decision Trees (GBDT))) and xgboost . were trained and analyzed.

1.SVM: SVM stands for Support Vector Machine. A support vector machine is a model for two-group classification problems using classification algorithms. SVM is a stable and fast algorithm. It consists of a line, called a decision boundary, that separates two data objects.

This serves as the pivot of separation. Separating axis equation:

Y = mx + c,

where m is the slope. The hyperplane equation separating the data entities is now H:wt(x)+b = 0. where b is the offset.

2. Decision tree (DT): A decision tree is a decision-making tool that contains probabilities of event outcomes, resource costs, and utilities. It is used for predictive modeling, statistics, and machine learning. Information gain is an important parameter in decision trees because it reduces the amount of information needed to differentiate between two data points for a partition.

$$Info(D) = \sum_{m} pi \log z\,(pi)$$

$$InfoA(D) = \sum_{vj=1} \frac{|DJ|}{|D|} * \mathrm{Info}(\mathrm{Dj})$$

In this case, pi represents the probability that a tuple in data set D is of class ci. Info(D) is the average amount of data required to determine which data object class D it belongs to. The info gain is calculated as

Gain(A) =info(D) – infoA(d).

3. **Random Forest**: A random forest is a classifier that uses multiple decision trees for different subsets of a given data set and averages them to improve the predictive accuracy of the data set. Rather than relying on a single decision tree, a random forest takes predictions from each tree and predicts the final outcome based on the majority of predictions. Scikit-learn assumes only two child nodes (binary tree) and uses Gini importance to calculate the importance of nodes in each decision tree.ni sub(j)= the importance of node j

w sub(j) = weighted number of samples reaching node j

C sub(j)= the impurity value of node j

left(j) = child node from left split on node j

right(j) = child node from right split on node j

4. Extra Tree Classifier: Extremely Randomized Trees Classifier(Extra Trees Classifier) is a type of ensemble learning technique that aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. In concept, it is very similar to a Random Forest Classifier and differs only in the way the decision trees in the forest are constructed. The Extra Trees Forest's

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

Decision Trees are built from the original training sample. Then, at each test node, each tree is given a random sample of k features from the feature set, from which each decision tree must choose the best feature to split the data according to some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple de-correlated decision trees

The analysis results are presented in Fig 4.4 and 4.5, and it can be seen that the Extra Tree Classifier is the most effective model. These classifiers were run using the scoring matrix parameters (precision, recall, and F-score). It was done in four different ways. True Positive (TP) = Depressive case that is positive and expected to be positive. A true negative (TN) case of depression is one that is negative and expected to be negative. False Negative (FN) depression cases are those that are positive but are expected to be negative. False Positive (FP) depression cases are those that are negative but are expected to be positive.

Precision is the ratio of true positives to cases that are expected to be positive. It is the level of selected cases that is correct.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positive}$$

Recall is the proportion of true positives to the cases that truly positive. It is the level of chosen cases that are selected

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Accuracy is one metric for evaluating classification models. Informally, **accuracy** is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Algorithm	Accuracy	Precision
ETC	0.93700	0.963631
RF	0.92975	0.951880
SVC	0.93775	0.941810
LR	0.93500	0.938043
NB	0.87025	0.865136
GBDT	0.85500	0.855118
DT	0.84700	0.847273
xgb	0.84775	0.846313
KN	0.84750	0.845201

Fig 4.4 Performance Metrics of ML Classifiers

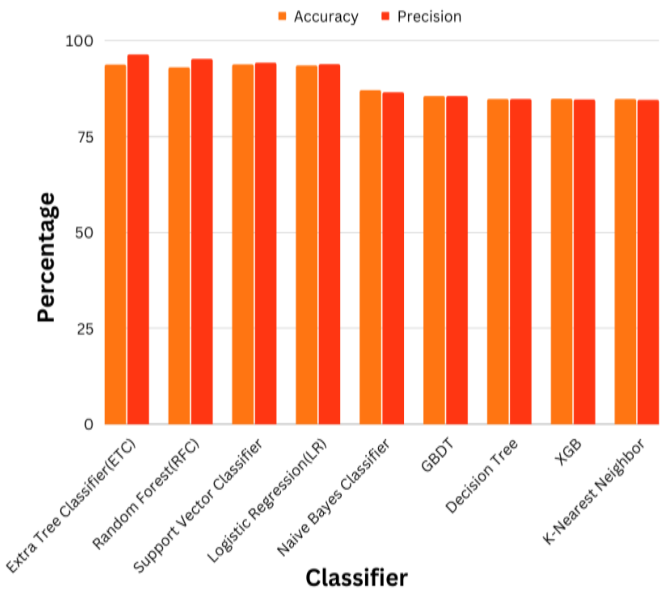


Fig 4.5 Visual Representation of Performance Metrics

It is clear that the Extra Tree classifier gives excellent results. We believe that this study paved the way for future research on inference and discovery of additional information based on causal events, such as discovery of hidden emotions or causes, prediction of public opinion based on causal events, etc.

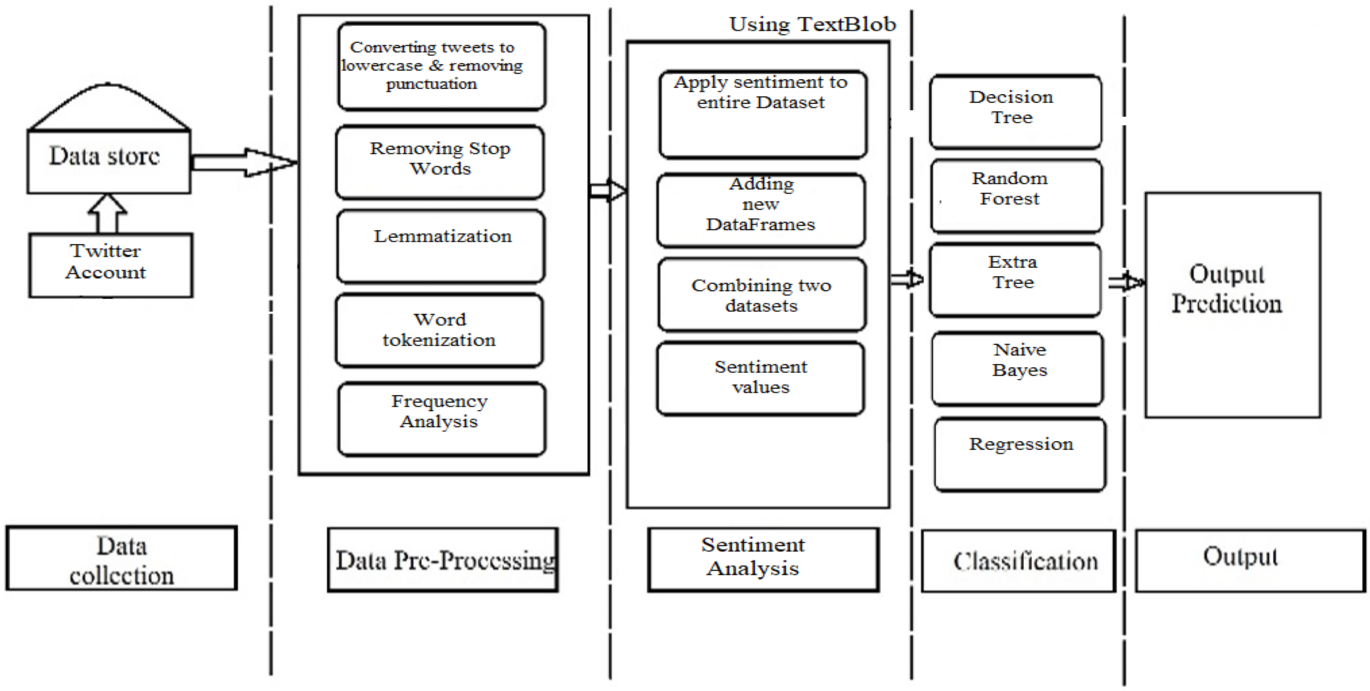


Fig 4.6 System Architecture

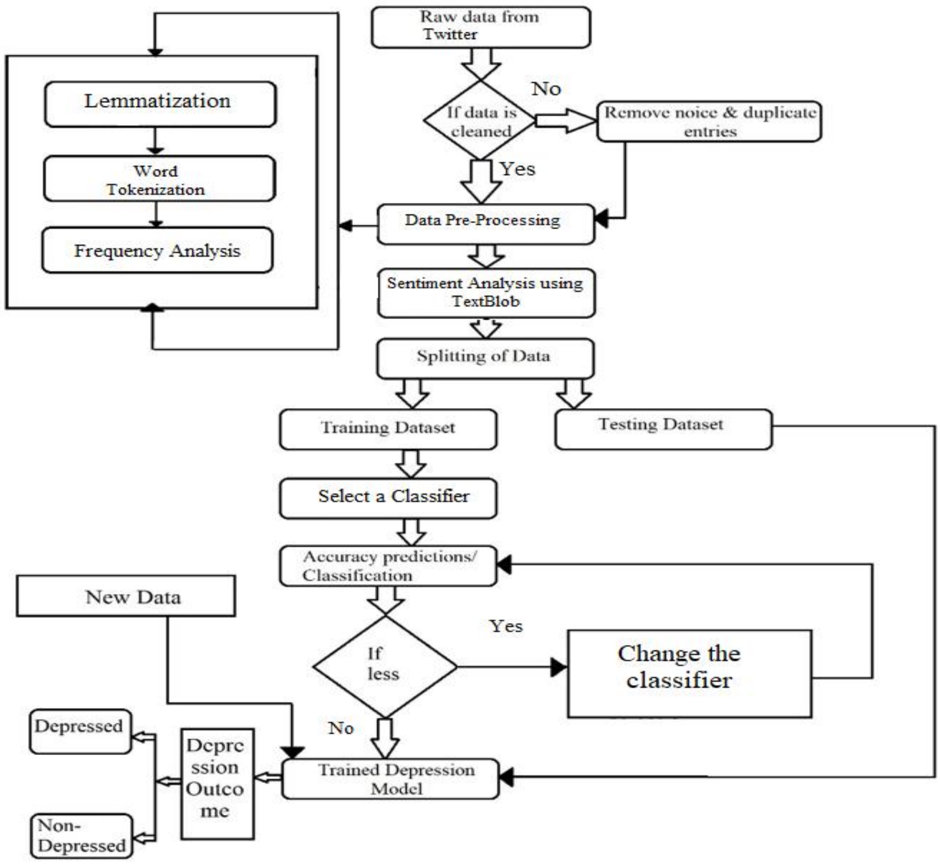


Fig 4.7 System Flowchart

5. Result and Conclusion

We have exhibited the capability of using Twitter as a tool for measuring and detecting major depression among its users. Worked on Twitter users' accounts for depressive behavioural exploration and detection. The data was collected from the social network. Depression was detected on various factors like followers, friends, post_text, post_created, post_id, favourites, requests.Sentiment analysis/Opinion Mining understands the feelings, replications as well as judgements amassed or extracted from texts or other data utilized in data analysis or mining, web mining, and convivial media analytics because sentiments are to judge human comportment

Then supervised machine learning approaches were applied to study each factor types independently. The classification techniques such as 'Decision tree', 'Random forest', 'Support Vector Machine', 'Extra tree classifier' ,'naïve bayes' and 'regression' are deemed suitable for each type. The result shows that in different experiments Extra tree classifier gives the highest accuracy than other Machine Learning approaches to find the depression as shown in the fig 4.5

6. Future Work

Depression can influence any of us anytime. However, some phases or events make us more vulnerable to depression. Physical and emotional changes associated with growing-up, losing a loved one, beginning a family, retirement may trigger some emotional infux that could lead toward depression for few people.

The are several different ways to relegate sentiments.

As of future work, various other data can withal be utilized for the depression detection. For example, biometrics data, Facial expressions of the user, Speech signals of the user and EEG signals. With the social media data, these data additionally was auxiliary for the analysis of the detection. The cumulation of different algorithms can additionally be habituated to check the precision value under different conditions and with different data. We could also increase our training data size by using various sampling methods.

Also we could introduce multiple languages in the future.

REFERENCES:

[1] A.G. Reece, A.J. Reagan, K.L. Lix, P.S. Dodds, C.M. Danforth, and E.J. Langer, "Forecasting the onset and course of mental illness with twitter data," Scientific repobrts, vol.7, no.1, p.13006, 2017

[2] Psychological Analysis for Depression Detection from Social Networking Sites. (Sonam Gupta , Lipika Goel , Arjun Singh , Ajay Prasad , and Mohammad Aman Ullah) 6 April 2022

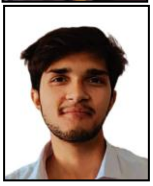
[3] Depression detection from social network data using machine learning techniques (Md. Rafqul Islam, Muhammad Ashad Kabir , Ashir Ahmed , Abu Raihan M. Kamal , Hua Wang4 and Anwaar Ulhaq) 27 August 2018

[4] Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features(Hatoon S. ALSAGRI, Mourad YKHLEF)24 April 2020

[5] Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence Intelligence.(Nirmal Varghese Babu,E. Grace Mary Kanag a)19 November 2021

BIOGRAPHIES

Amaan Khan
Third Year Engineering
Computer Engineering
Rajiv Gandhi Institute Technology



Jayeshkumar Kharol
Third Year Engineering
Computer Engineering
Rajiv Gandhi Institute Technology

Abhishek Gaikwad
Third Year Engineering
Computer Engineering
Rajiv Gandhi Institute Technology

Swapnil Jagtap
Third Year Engineering
Computer Engineering
Rajiv Gandhi Institute Technology

Prof. Suresh Mestry
Assistant Professor
Computer Engineering
Rajiv Gandhi Institute Technology