Predicting Depression in Social Network Sites Using NLP

BY

Md. Tazmim Hossain
ID: 173-15-10390
AND
Md. Arafat Rahman Talukder
ID: 173-15-10404

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science & Engineering

Supervised By

Nusrat Jahan
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Gazi Zahirul Islam
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATION UNIVERSITY

DHAKA, BANGLADESH

SEPTEMBER 2021

APPROVAL

This Project/internship titled **"Predicting Depression in Social Network Sites Using NLP"**, submitted by Md. Tazmim Hossain, Md. Arafat Rahman Talukder ID No: 173-15-10390, 173-15-10404 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 9th September,2021.

BOARD OF EXAMINERS

Chairman

Dr. TouhidBhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Nazmun Nessa Moon
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Fizar Ahmed
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Md Arshad Ali
Associate Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology
University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Nusrat Jahan, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Nusrat Jahan

Nusrat Jahan
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Zahir

Mr. Gazi Zahirul IslamLecturer
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

MD. Tazmim Hossain

Md. Tazmim Hossain

ID: 173-15-10390
Department of CSE
Daffodil International University

Md. Arafat Rahman Talukder

____

Md. Arafat Rahman Talukder
ID: 173-15-10404
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

At First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year thesis successfully.

We really grateful and wish our profound our indebtedness to **Nusrat Jahan**, **Lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Natural Language and Processing*" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Touhid Bhuiyan, Professor and Head**,** Department of CSE, Faculty of Science and Information Technology, DIU, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

Atlast,againwewanttothankallthegoodwishers,friends,family,seniorsforallthehelpandinspirations.Thisresearchisaresultofhardworkandallthoseinspirationsandassistance.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Depression is an acute problem throughout the world, where more than 264 million people are suffering from it. Due to worst and prolong depression near around 800000 people dies in every year. The real problem is that most of the people are not concern of the fact that they are suffering from depression. Here, our aim was to find out whether an individual is in depression or not by analyzing social media text information. Our dataset consists of 1500 sentences, which was collected from different social media platforms– Facebook, Tweeter, and Instagram. Then we have performed some data preprocessing approaches such as– tokenization, remove of stop words, remove of empty string, remove of punctuations, stemming and lemmatizing. After data preprocessing, we considered processed text as input. We work on six different machine learning classifiers which produced great accuracy over our dataset. Among six algorithms, Multinomial Naive Bayes and Logistic Regression provided 95% accuracy.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Introduction:

Depression is a usual and significant medical disease which negatively affects how a person feels, how he acts and the way he thinks. It is the key reason of inability worldwide. People of all ages can suffer from depression. People are losing their ability to do their daily work due to depression and this is pushing them towards more severe depression. The biggest problem is that the people who are depressed do not know that they are suffering from depression. That is why they cannot help themselves. As a result, they are being gradually suffered severe mental and physical damage. After suffering from depression for a long time, they think that they have no choice but to commit suicide. We always wanted to do something for the community and finally we got the idea of dealing with depression. In this global pandemic people are getting more depressed than other times. It is becoming a major issue day by day. To reduce the suicidal case, we tried to build a prototype on which depression can be identified. In our research work we have applied supervised machine learning technique. Our dataset is labeled with two heading named as depressed and non-depressed. We have collected data from different social media platforms such as Facebook, Tweeter and Instagram from 200 users. By implementing of Vader we found out the initial accuracy but which was far below our expected output . After that we have cleaned and preprocessed our dataset and applied seven different machine learning classifiers and among all of them we got the highest accuracy of 95% from Naïve Bayes. First of all we had worked with 1500 data after that we had enhanced our data to 2000. Therefore, our accuracy also increased from 95% to 98%.

.

1.2 Motivation:

From the very beginning we were very interested and excited about our research project. We wanted to do something different. That's why we have been dealing with different research areas for a long time. Then we decided that we will work on machine learning and NLP to detect depression from social media. The Covid-19 epidemic has increased people's frustration and is having a devastating effect on their daily lives. Itis an ecumenical problem all-over the world, where more than 264 million people are suffering from it [WHO]. Due to worst and prolong depression near around 800000 people dies in every year [WHO]. People are losing their ability to do their daily work due to prolonged depression and they feel so much loneliness that their willpower to survive is declining and they are finally choosing the path of suicide. This is the main reason that has inspired our work a lot. Our main goal is to safe the entire mankind from being depressed and to detect depression in a computerized way .

1.3 Rational of the study:

People express their thoughts and emotions through various social media such as Facebook, Twitter and Instagram and these emotions are usually expressed through videos, pictures and text. Textual data has become a widely used and popular medium of communication.

Machine learning classifiers are very useful of predicting whether textual information is presenting depressed information or Non-depressed information. Machine learning is classified into two sections known as unsupervised learning and another is supervised learning. In machine learning supervised learning is a technique where we train the machine from a dataset where some data is already tagged with correct answer or the dataset is well labeled. Moreover, in unsupervised learning the dataset is not labeled or classified therefore the machine has to find out the pattern, similarities and differences from an unsorted data to apply algorithms. In our research work we have applied supervised machine learning technique.

1.4 Research questions:

It was very difficult for us to get the job done because we collected users' personal text data with their permission. Therefore, it cost a lot of time. Then the major challenge was to bring the expected accuracy. We faced many questions about all this. Below they are given:

1. How can we collect data?
2. How to sort the dataset?
3. How can we preprocess our dataset?
4. How can we know which algorithm will work best for our dataset?
5. How to bring best accuracy?

1.5 Expected Output:

Our main objective was to find out how to find people's depression from textual data. With the help of various machine learning techniques and algorithms we were able to find it and its accuracy was 98%.

1. Applying supervised based classification we have been able to detect depression using textual data.
2. First of all we had got the accuracy of 95% after that we had increased our dataset and our accuracy also increased to 98%
3. Through this, people will be able to find out their depression very easily.
4. Therefore it will free them from the tendency to commit suicide

CHAPTER2: BACKGROUND

2.1 Terminologies:

Related works, Comparative analysis and summary, scope of the problem and challenges has been discussed in this section. In related work section we have deal with the research papers that are related to our research work. We discussed how they collected datasets, how they processed data, what algorithms they used, and their accuracy rate. In Comparative Analysis and Summary, we tried to find out the appropriate classifiers and methods for our work. We have analyzed over lot of ways to increase the accuracy level. We labeled up our dataset in most efficient way also works on reducing complexity. For summarizations we've finalized the ultimate classifiers which provide an accuracy of 98%. In Scope of the problem we have discussed the problems we faced while preprocessing, cleaning and applying classifiers on our dataset. In challenges section we have discussed the difficulties we have faced while collecting our dataset and also represent it in a machine-readable form.

2.2 Related work:

Mandar Deshpande(2017) used Twitters API for data collection. For pre-processing data, Count-vectorizer, tokenizing words, stop word removals, POS tagging have used. 20% of data was used for testing purpose and 80% of data was used for training purpose. To classify the data Word2vector was used here.At first SVM was applied its accuracy was 79% but when Naïve bayes was applied its accuracy was 83%. Supervised learning was used in this research paper. [1]

Priyanka Arora(2019) and Parul Arora (2019) applied Support Vector Regression & Multinomial Naïve Bayes with different types of approach and compared their accuracy with other authors. SVR & Multinomial Naïve Bayes both were applied as classifier SVR performed better than Multinomial Naïve Bayes and its accuracy was 79.7%. Twitter API was used for collecting data. Tokenization, Stemming, Gram features, Sentiment extractions, POS vector was applied for feature extraction. Their accuracy rate was increased than others related work. [2]

Nafiz Al Asad(2019) used Facebook, Twitter & Question based evaluation for data collection. Beautiful Soup was

applied for data preprocessing from Twitter & also JSON data was converted to CSV format for Facebook data collection. For pre-processing data NLTK library was used. They divided depression label into 6 categories they are- Normal, Mild, Borderline, Moderate, Severe, Extreme. SVM & Naïve Bayes both were applied as classifier but Naïve Bayes performed better than SVM and its accuracy was 74%. [3]

Kantinee Katchapakirin(2018) selected Thai language to determine depression using microblogging sites and questionnaires. NLTK & Weka both were applied for data preprocessing. As SVM & Random Forest were used as classifier and also a Deep Learning model was applied. Among all of 3 three model Deep Learning performed better and its accuracy rate was 85. Only 35 Facebook users post was fed for training and testing purpose Also due to the Thai language some words couldn't be converted into English directly as a result polarity score came out poor. [4]

S.A.S.A. Kulasinghe (2019) built a chat-bot to having a conversation with users and based on that conversation data was collected from users. Facebook status was also used as a data collection process. Along with text voice was also used for data collection. NLTK and Textblob python library was applied for text pre-processing. 80% data was kept for training purpose and 20% data was kept for testing purpose. Many machine learning classifiers was applied but among all of them SVM performed much better than the others and final accuracy rate was 95%. [5]

Rashedul Amin Tuhin(2019) worked on Bengali language to find out emotion from text. Data set was created manually and 7500 sentences was collected for the corpus. Data pre-processing methods were not much appropriate and efficient. Naïve Bayes and Topical Approach were applied as classifiers but Topical Approach performed better than Naïve Bayes and its accuracy rate was more than 90%. For better performance they could use LDA.[6]

Nusrath Tabassum (2019) have worked on Bengali language. Dataset was collected from Facebook & Twitter but data collection was comparatively small only 1050 text was collected. Data pre-processing have done by using NLTK library & tokenization. Only one classifier Random Forest was applied. Here, they could have used more classifier in order to have better accuracy. [7]

Nushrat Jahan Ria(2020) proposed 6 model to identify Saint & Common form in Bengali text. 1200 mix Bengali sentence (Common & Saint) was used as dataset. Data pre-processing wasn't much appropriate as they didn't applied lemmatizing, Steaming, bag of words process. They only removed Stop Words. They used only 1200 mixed sentences but if they used more than 2000 sentences then their accuracy may also increase. 80% data was applied for training purpose & 20% data was applied for testing purpose. Though They have used six classifiers among all of them Naïve Bayes accuracy came out highest which is 77%. Accuracy rate could be improved if they have taken proper initiative in data collection & pre-processing. [8]

Sandeep Nigam(2018) proposed a machine learning based approach on Sentimental Analysis. Sentiment140(Stanford University) was used as dataset. NLTk, Beautiful Soap, TF-idf, CountVectorizer was applied for data pre-processing. With stop words (SW) & without stop words both were applied among Unigram, Bigram, Trigram but with stop words accuracy was better than without stop words & among them Trigram accuracy rate was better. They have used different machine learning algorithms but among of all them Logistic Regression performed better &it's accuracy rate was 82.59%. Depend on the comparison, Tfidf with stop words gives more accurate result than CountVectirizer. [9]

Md. RakibulHasan(2019) have developed a self-model to predict sentiments on products based on twitter API. For Data pre-processing NLTK, Stemming, Lemmatizing, Bag of words, Stop words removal, POS tagging, Entity recognition were applied. Feature Extracted from Twitter data &Tf-idf was applied to find out the most frequency words. Including self-developed method, they have applied different types of classifier such as SVM, Naïve Bayes, Maximum entropy, k-nearest neighbor & among all of them their self-developed model performed better which accuracy was 82.25%. [10]

VikasGoel(2018) worked on Multilingual language based on Machine learning & Deep Learning approach. Twitter was used for data collection. Google API was used as they worked on multilingual language. Though they have used Google API as translator but Google API can't translate deep meaning of a sentence properly. Supervised learning was applied for this work. For pre-processing they have removed URL's, Stop words, Slangs, Misspelled words, re-tweets etc. Online Stemmer was applied for feature extraction. Then the feature was expressed in numerical data to apply on classifier. Two types of classifier were used they are- RNN & Naïve Bayes. Among of them RNN performed better which accuracy was 96%. [11]

Enrico Laoh (2019) worked on hotel review to find out the sentiment of a review whether it is positive or negative. Data was collected from "Tripadvisor.com" & location was Bali, Indonesia. Split sentence & Tokenization was used for data pre-processing. They didn't remove Stop words. SVM was applied as classifier & accuracy was 94%. To increase the level of accuracy unigram & Bigram were used among of them Bigram performed better with accuracy 94%. For the same data set they have applied RNTN classifier but the accuracy was 85% which was lower than SVM. They could have completed the data pre-processing phase more efficiently by using CountVectorizer&tfidf-Vectorizer to increase the accuracy score. [12]

Anees Ul Hassan(2017) worked on sentence level sentiment analysis for depression measurement. They have made a comparison among three classifiers & they are SVM, Naïve Bayes, Maximum Entropy. Data was collected from Twitter & 20 news group. For data pre-processing at first, they have split the sentence then tokenizing the words, removal stop words & steaming was applied. For feature extraction N-gram, POS tagging, Negation, Sentiment Analyzer, bag of words was applied. Meta learning(voting) was used for comparing above three classifiers. Among three classifiers SVM performed better & it's accuracy was 91%, NB accuracy was 83% & ME accuracy was 80%.[13]

Dipti Mahajan(2018) worked on sentimental analysis using RNN & Google Translator. Data was collected from Twitter API &it's size was 10,000 sentences. They have used Google Translator API to convert non-English language to English language. For data pre-processing pointless words were removed and for increasing the accuracy StandfordNLP library was applied. Three types of classifier were used they are RNN, Naïve Bayes, SVM and among all of three RNN performed better which accuracy was 90.3%. [14]

Dilesh Tanna(2020) build a model to analyze user sentiment. They have created a social media platform where user can like, comment, post & share. Here, users are defined as happy or sad based on their social media activities such as if a user post something negative then it'll decrease the rating score of that particular user and if a user comment something positive on a post then the rating score of that user will be increased. This rating scores are stored in backend Database which is only accessible by the Admin panel. Three types of classifier were used in that model which are SVM, Naïve Bayes & Maximum Entropy. These classifier helps in processing the analysis more effectively. [15]

2.3 Comparative analysis and summary:

At first, we have implemented VADER and tried to find out the initial accuracy but it was far away from our expected outcome therefore it's accuracy was 85%. Usually VADER returns the polarity of a sentence whether it's +ve,-ve or ve.After that we have preprocessed our dataset by removing punctuations, stop words, empty string and applied lemmatization, stemming, tokenization to convert the dataset into machine readable form. After implementing VADER we applied six above classifiers on our dataset to find out the accuracy which provides NB=95%, LR=95%, RF=93%, KNN=91% and DT=89%. Among the six classifiers NB & LR provide the highest accuracy of 98%. After that we had increased our dataset from 1500 data to 2000 data and our accuracy had also increased.

For 1500 data:

Table 2.3.1: Accuracy for 1500 data

Classifier	Precision	Recall	F1-score	Accuracy
Multinomial NB	0.95	0.95	0.95	95%
LR	0.95	0.95	0.95	95%
Linear SVC	0.94	0.94	0.94	94%
RF	0.93	0.93	0.93	93%
KNN	0.91	0.90	0.91	91%
DT	0.89	0.89	0.89	89%

For 2000 data:

Table 2.3.2: Accuracy for 2000 data

Classifier	Precision	Recall	F1-score	Accuracy
Multinomial NB	0.98	0.98	0.98	98%
LR	0.96	0.97	0.97	97%
Linear SVC	0.95	0.97	0.96	96%
RF	0.93	0.92	0.93	93%
KNN	0.95	0.84	0.89	91%
DT	0.92	0.86	0.89	90%

2.4 Scope of the problem:

We've faced couple of problems and scope of these problem is given below-

1. We couldn't find out the appropriate strategy to start our research work then we divide each individual part into sub category that helped us to implement this work faster.

2. We've faced big trouble in data preprocessing section. Because our data was web based so it was very complicated to extract in normal form. But we've used diffrent preprocessing techniques such as StopWord removal, Null value elimination, Removing punctuation & white spaces which helped us to make simplify the dataset.

3. We were very tensed about data collection then we fix this issue by collecting real-life data from various social media platform.

4. As we applied vader to check the polarity of each sentence and to identify the primary accuracy level but it provides us a very less number of accuracy. Then we resolve the problem by implementing different supervised learning classifier which provide us the maximum level of accuracy.

2.5 Challenges:

The major challenge was to collect the dataset. As we collect data from different micro blogging sites so the format of the data was different. So we had to convert all the data in one format which was quite challenging for us. As our dataset was very large so it took a lot of time to provide the output. We had to take the users permission to collect their textual information which was very tough for us. We had to remove the null values, empty string, punctuations, stop words, emojies and links from our dataset and took a bit of time . Then finding the classifiers was also a challenging part for us. Therefore we search for different classifiers which will provide us the highest accuracy.

CHAPTER3: RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation:

Our research topic is “Depression Analysis using Natural Language Processing (NLP)”. Before selecting our topic we have spent a lot of time to find our interested field. Finally, we had found our interest on Natural Language processing (NLP). Due to covid-19 pandemic people are getting depressed day by day and we thought that it is high time to analysis this problem and detect depression. We've used Jupyter Notebook which is basically use for Python development, Microsoft Windows platform, Multiple Editors for designing, various Python advance libraries such as Numpy, Pandas, Matplotlib, Seaborn, Cufflinks , NLTK, regular expression package, scikit-learn, wordcloud etc to develop our project .

3.2 Data collection procedure / Data set utilized:

We have collected data from different social media platforms such as Facebook, Tweeter and Instagram from 200 users. At first we couldn't find proper resources to collect data. As we wanted to have real life user data so we fixed our mind to collect data from different social media because people love to express their thoughts on social media. We had to take users permission as these data collected from users account. Some people denied providing their data as a result we had to reach more people for collecting data. Then again, we had to modify these data into general form as our collected data was in complex web format. After that we converted data to numerical format to apply in classifier.

3.3 Statistical analysis:

In this research work we have applied supervised learning method. The working procedure is illustrating step by step in the below section.

A. Collection and Properties of Dataset:

Our dataset is consisting of total 2000 sentences where 1023 were tagged as Non-Depressed sentences and 975 were tagged as Depressed sentences.

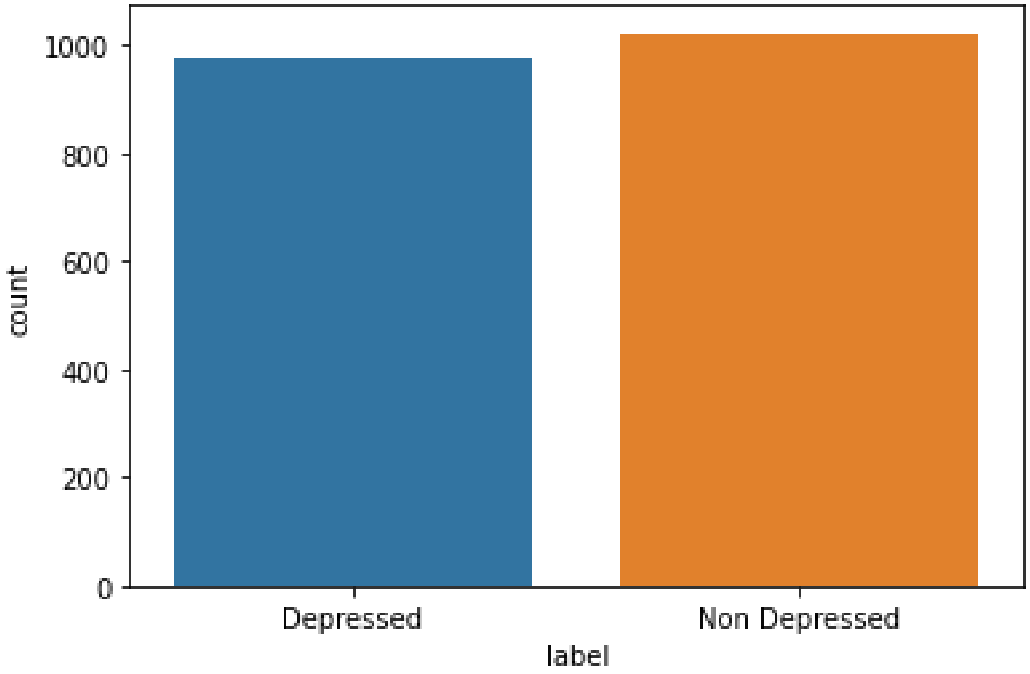
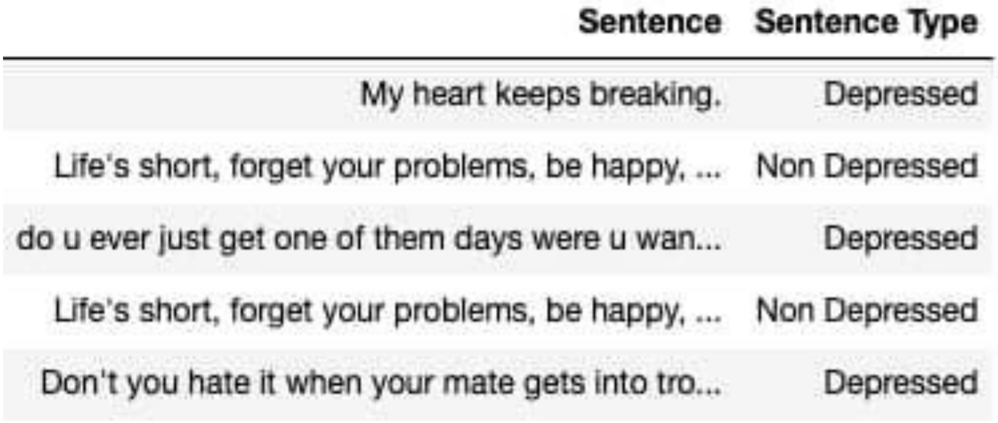


Figure 3.3.1: Ratio of Depressed and Non-Depressed data

Our data set is consist of 2000 sentences where 1200 sentences were collected from 200 Facebook users, 500 sentences were collected from 300 tweeters users and 300 sentences were collected from 100 Instagram users.



Sentence	Sentence Type
My heart keeps breaking.	Depressed
Life's short, forget your problems, be happy, ...	Non Depressed
do u ever just get one of them days were u wan...	Depressed
Life's short, forget your problems, be happy, ...	Non Depressed
Don't you hate it when your mate gets into tro...	Depressed

Figure 3.3.2: Head of dataset

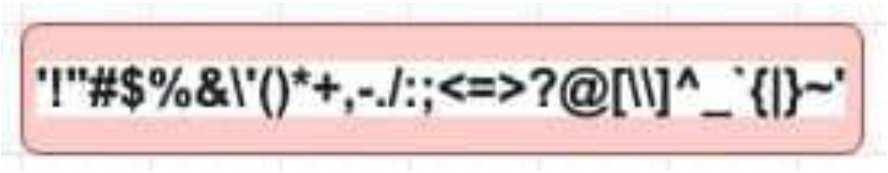
B. DATA SET DISTRIBUTION

Among 2000 sentences in our dataset 1600 sentences were used for training purpose and 400 sentences were used for testing purpose. There are 200Non-Depressed sentences and 200 Depressed sentences which we have tested.

c. Data Preprocessing

The EXCEL file is read and different data preprocessing steps are applied on it. The applied preprocessing steps are given below:

1. Data normalization: After importing our dataset we have checked whether there is any null input or not. We have found out two null inputs in our dataset then we have dropped those null inputs. All types of punctuations have been removed from our dataset as they don't put any impact in our research work. List of punctuations that we removed from our dataset are:
2.



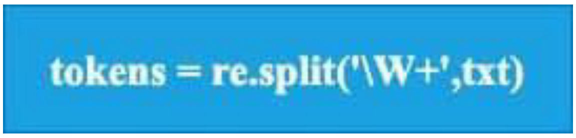
```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

Figure 3.3.3: Removing Punctuation

Example: What a sunny day !!! I would love to go outside. What about you?

After removing punctuations: What a sunny day I would love to go outside What about you

2.Data Tokenization: Tokenization is the method of splitting or tokenizing a string. Words are the tokens of sentences and the sentences are the tokens of paragraphs. We have applied sentence to word tokenization for our dataset. We have used python split function to tokenize our sentences in words.



```
tokens = re.split('\W+',txt)
```

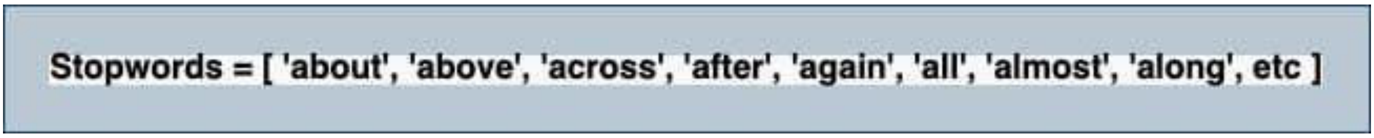
Figure 3.3.4: Splitting sentences

Example:

Table 3.3.1: Before and After Tokenization

Before tokenize	After tokenize
My heart keeps breaking.	[my, heart, keeps, breaking]
Life's short, forget your problems, be happy	[lifes, short, forget, your, problems, be, happy]
I am feeling very lonely.	[I ,am,feeling,very,lonely]
Don't you hate it when your mate gets into	[dont, you, hate, it, when, your, mate, gets, intro]

3. Remove of stop words: The most used words are known as stop words.



```
Stopwords = [ 'about', 'above', 'across', 'after', 'again', 'all', 'almost', 'along', etc ]
```

Figure 3.3.5: Removing Stopwords

They don’t have any significant to identify depression level from a sentence. Therefore, we have removed those from our dataset. We have built our own stop words considering the fact that they don’t change the meaning of a sentence.

Table 3.3.2: Before and After Removing Stop Words

Before Removing stop words	After removing stop words
This is a beautiful country and I am very glad to see that.	beautiful country very glad see

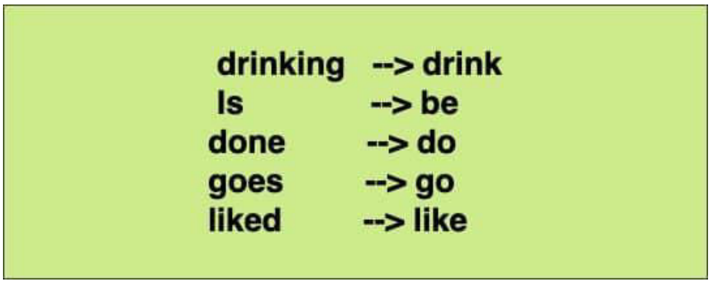
4. Empty string remove: Empty string can be very sensitive while implementing the classifiers. It kills our memory space and we may get lower accuracy rate. Therefore, we have removed all empty string from our dataset.

Table 3.3.3: Before and After Removing Empty String

Before removing empty string	After removing empty string

[“ You”, “ ”,“ are”, “ ”, “looking”,” ”, “ gorgeous”]	[“You”,”are”,”looking”,”gorgeous”]

5. Lemmatization: Lemmatization is the method of removing inflectional endings from a particular word and it returns the base or the dictionary form of that word.



drinking	--> drink
Is	--> be
done	--> do
goes	--> go
liked	--> like

Figure 3.3.5: Lemmatization

Text preprocessing helps to increase the accuracy of the classifiers. We have performed the below text preprocessing techniques to increase the accuracy rate of our classifiers.

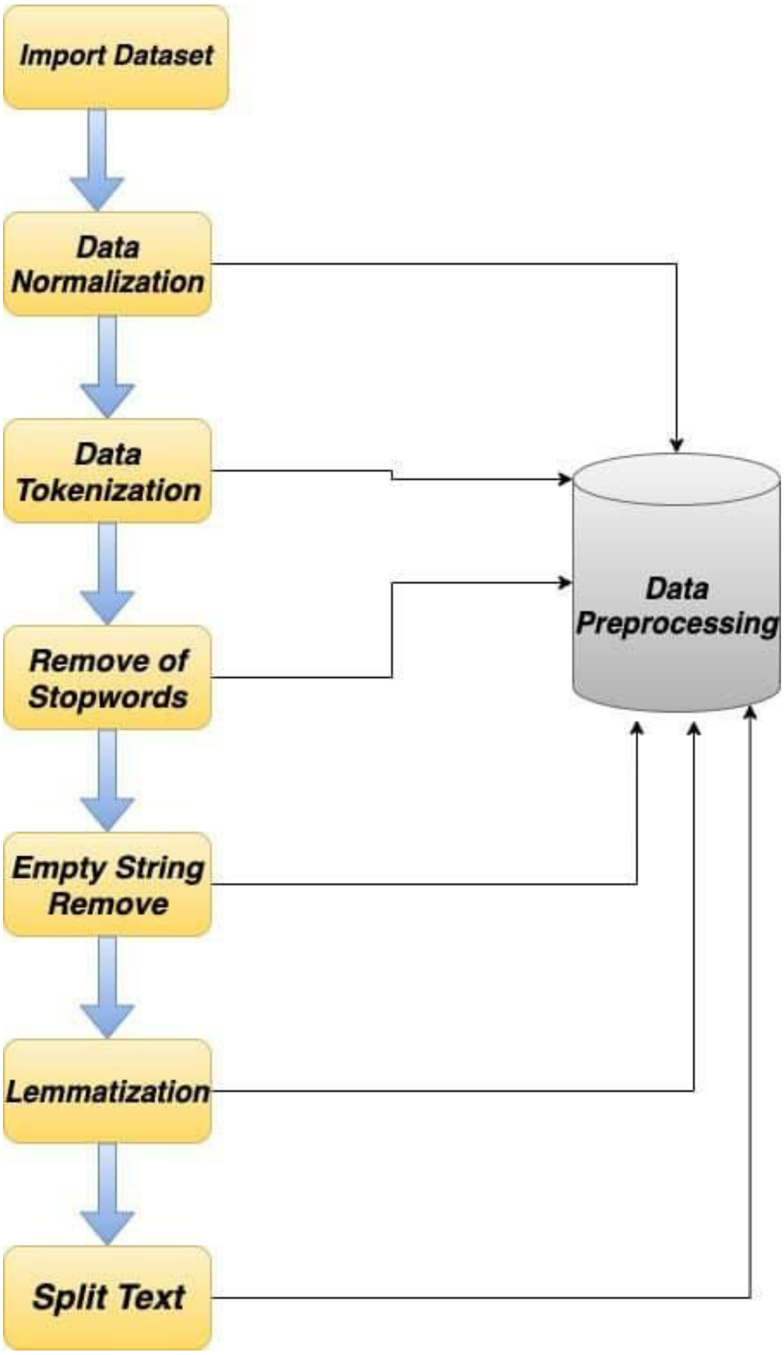


Figure 3.3.6: Data Pre-processing

3.4 Proposed Methodology/Applied mechanism:

We have created a model using machine learning techniques on how to analyze human depression through text. Among three broad areas of machine learning we have used supervised learning approach because of its compatibility in designing and controlling dynamic processes. Nowadays, people express their feelings through social media and most of which are in text format. Vader was implemented initially to analyze the polarity of each sentence but the output was very poor. Then seven well known machine learning classifiers Multinomial Naïve Bayes, Linear SVC, KNN, RF, DT, SVM and LR were applied. Based on our success in our model, each classifier has a brief discussion below.

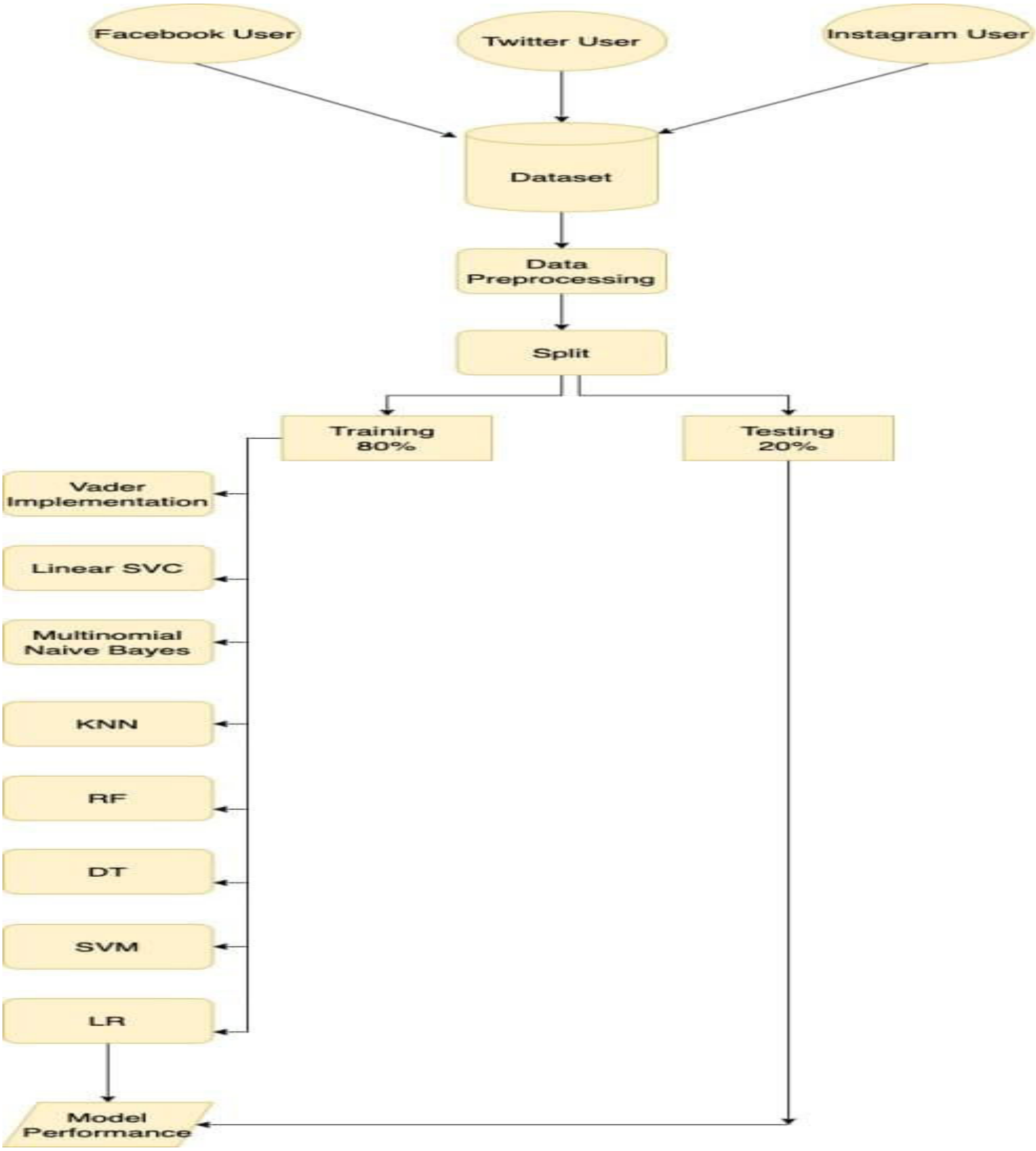
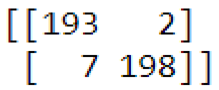


Figure 3.3.7: Working Procedure

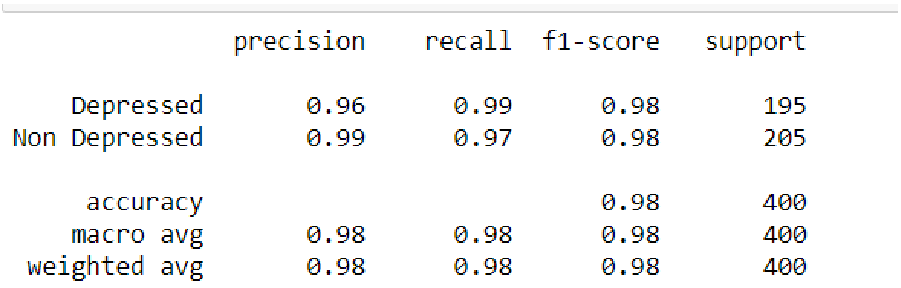
A. Naïve Bayes classifier: NB classifiers are the collection of different algorithms, It is not a single algorithm. In NB classifiers all the algorithms share their common principle. It is based on Bayes Theorem. For our dataset Multinomial Naïve Bayes performed very efficiently its confusion matrix was



```
[[193   2]
 [  7 198]]
```

Figure 3.3.8: Confusion Matrix of Naïve Bayes

Our classification report successfully predicted the output with 98% of accuracy.F1-score, precision, recall is very near to our output .



```
               precision    recall  f1-score   support

    Depressed       0.96      0.99      0.98       195
Non Depressed       0.99      0.97      0.98       205

     accuracy                           0.98       400
    macro avg       0.98      0.98      0.98       400
 weighted avg       0.98      0.98      0.98       400
```

Figure 3.3.9: Accuracy score of Naïve Bayes

Below figure is the basic rule of NB classifier:

$$P(A/B) = \frac{P(B/A)\ P(A)}{P(B)}$$

----------------------------(1)

Figure 3.3.10: Formula of Naïve Bayes

Linear SVC :

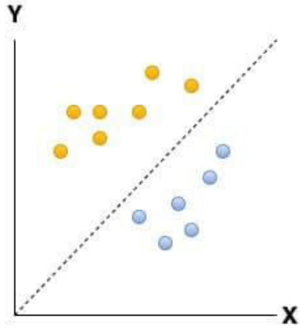
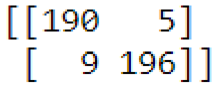


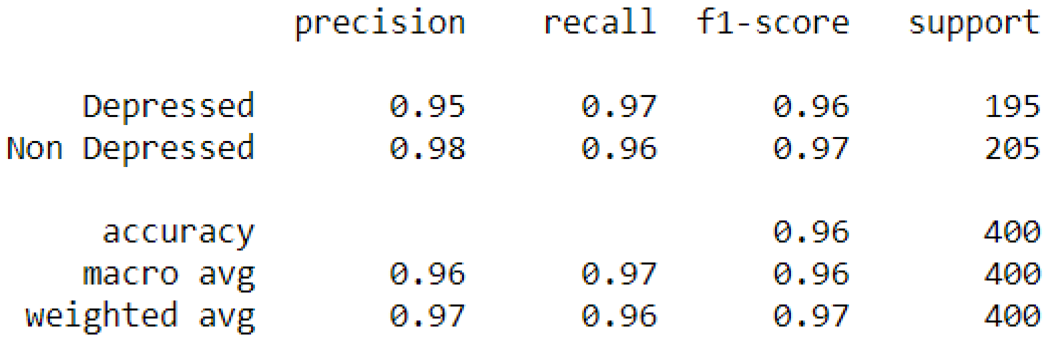
Figure 3.3.11: Diagram of Linear SVC

It applies a linear kernel function to perform classification and it is implemented in terms of liblinear. For N dimensional regression problems, it is used widely. On our dataset it gained the accuracy of 96%. Confusion matrix for Linear SVC is:



```
[[190   5]
 [  9 196]]
```

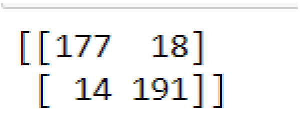
Figure 3.3.12: Confusion Matrix of Linear SVC



```
               precision    recall  f1-score   support

    Depressed       0.95      0.97      0.96       195
Non Depressed       0.98      0.96      0.97       205

     accuracy                           0.96       400
    macro avg       0.96      0.97      0.96       400
 weighted avg       0.97      0.96      0.97       400
```

Figure 3.3.13: Accuracy of Linear SVC

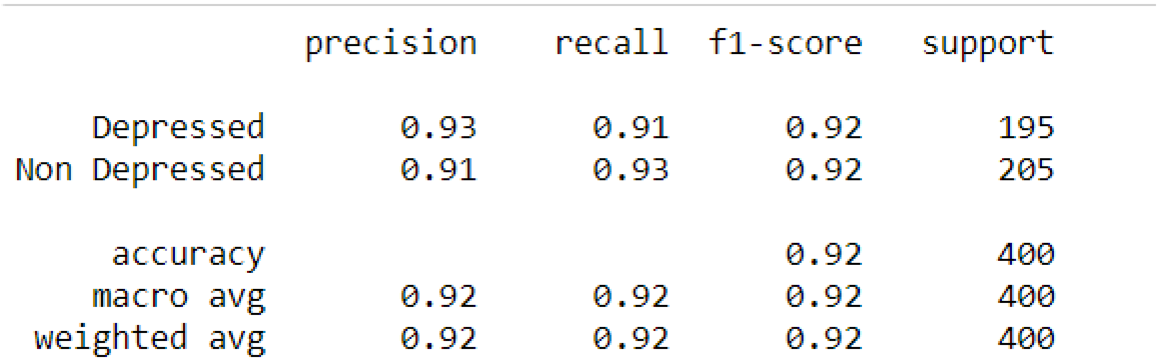
KNN classifier: KNN is a supervised learning algorithm. It is very simple and easy to implement. It is widely used to solve classification and regression problems. It performed very well on our dataset with the accuracy of 92%. Confusion matrix for KNN on our datsetwas :



```
[[177  18]
 [ 14 191]]
```

Figure 3.3.14: Confusion matrix of KNN

Related F1-score, precision, recall is very near to our output .



```
              precision    recall  f1-score   support

   Depressed       0.93      0.91      0.92       195
Non Depressed       0.91      0.93      0.92       205

    accuracy                           0.92       400
   macro avg       0.92      0.92      0.92       400
weighted avg       0.92      0.92      0.92       400
```

Figure 3.3.15: Accuracy of KNN

RF classifier :

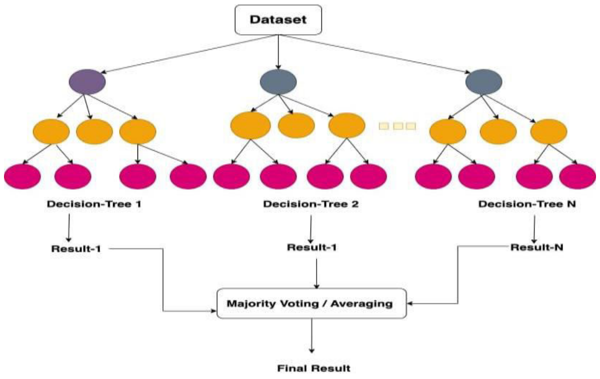
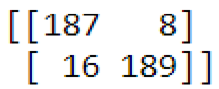


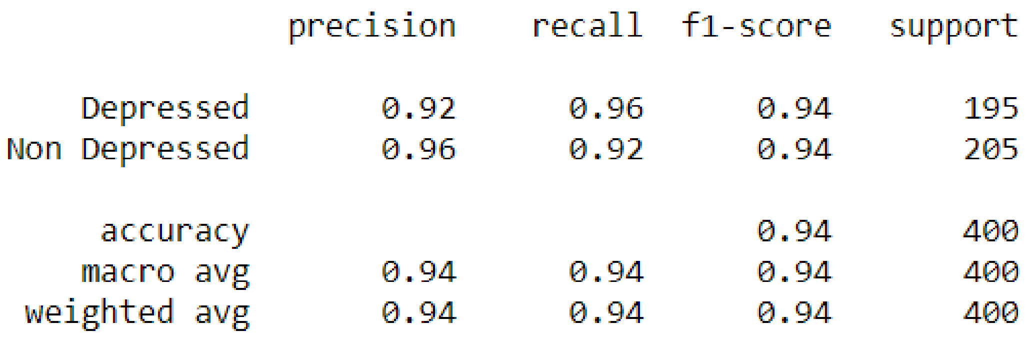
Figure 3.3.16: Diagram of Random Forest

RF is mainly used for classification problems. It is a supervised learning algorithm.IT Creates a decision tree based on data sample and by getting predictions from each of them select the optimal solution by voting. It ‘s accuracy on our dataset was 94% and it’s confusion matrix is:



```
[[187   8]
 [ 16 189]]
```

Figure 3.3.17: Confusion matrix of Random Forest



```
               precision    recall  f1-score   support

    Depressed       0.92      0.96      0.94       195
Non Depressed       0.96      0.92      0.94       205

     accuracy                           0.94       400
    macro avg       0.94      0.94      0.94       400
 weighted avg       0.94      0.94      0.94       400
```

Figure 3.3.18: Accuracy of Random Forest

Decision Tree:

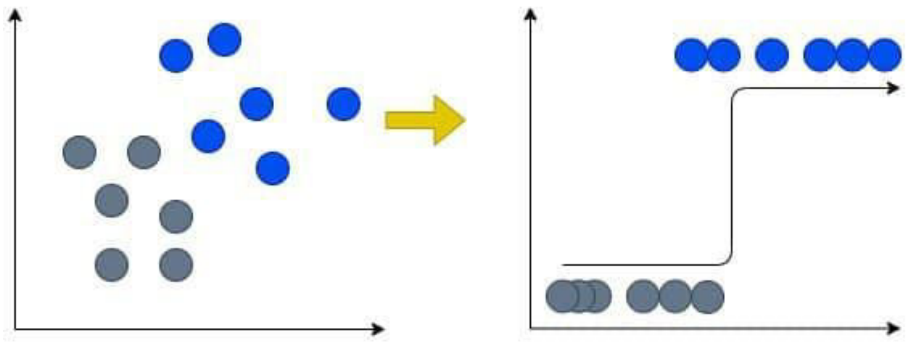
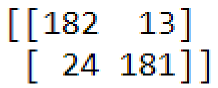


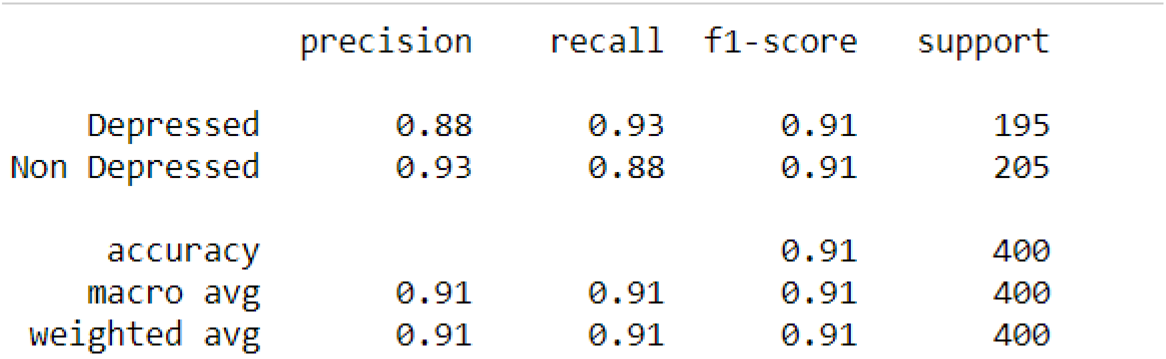
Figure 3.3.19: Diagram of Decision Tree

For classification and predictions DT is the most well-build and popular algorithm. It is a supervised classifier. It is widely used for classification problems but in some cases, it is also used for regression based problems. Comparing to other classifier it provides low accuracy of 91% . Confusion matrix of DT algorithm for our data set is



```
[[182  13]
 [ 24 181]]
```

Figure 3.3.20: Confusion matrix of Decision Tree

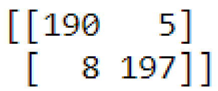


```
               precision    recall  f1-score   support

    Depressed       0.88      0.93      0.91       195
Non Depressed       0.93      0.88      0.91       205

     accuracy                           0.91       400
    macro avg       0.91      0.91      0.91       400
 weighted avg       0.91      0.91      0.91       400
```

Figure 3.3.21: Accuracy of Decision Tree

Logistic Regression:

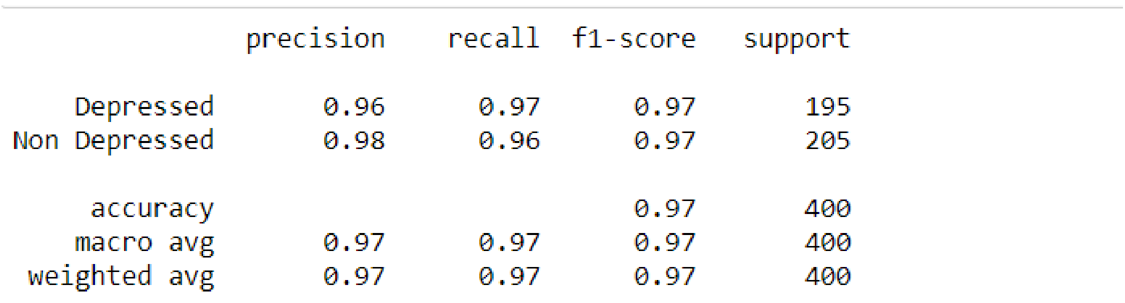
Generally, it is known as supervised learning classification algorithm. It is relatively fast and efficient comparing to other classification algorithms. It performed exquisitely on our dataset which provide an accuracy of 97%. Therefore, it’s confusion matrix is



```
[[190   5]
 [  8 197]]
```

Figure 3.3.22: Confusion matrix of Logistic Regression

Related F1-score, precision, recall is very closed to our output.



```
                precision    recall  f1-score   support

     Depressed       0.96      0.97      0.97       195
 Non Depressed       0.98      0.96      0.97       205

      accuracy                           0.97       400
     macro avg       0.97      0.97      0.97       400
  weighted avg       0.97      0.97      0.97       400
```

Figure 3.3.23: Accuracy of Logistic Regression

Experiment and output:

Table 3.3.4: Accuracy

Classifier	Precision	Recall	F1-score	Accuracy
Multinomial NB	0.98	0.98	0.98	98%
LR	0.96	0.97	0.97	97%
Linear SVC	0.95	0.97	0.96	96%
RF	0.93	0.92	0.93	93%
KNN	0.95	0.84	0.89	91%
DT	0.92	0.86	0.89	90%

By implementing VADER we tried to find out the initial accuracy but it was far away from our expected outcome therefore it’s accuracy was 85%. Usually VADER returns the polarity of a sentence whether it's +ve,-ve or ve. After that we have preprocessed our dataset by removing punctuations, stop words, empty string and applied lemmatization, stemming, tokenization to convert the dataset into machine readable form.

After implementing VADER we applied six above classifiers in our dataset to find out the accuracy which provides NB=95%, LR=95%, RF=93%, KNN=91% and DT=89%. Among the six classifiers NB & LR provide the highest accuracy of 95%.

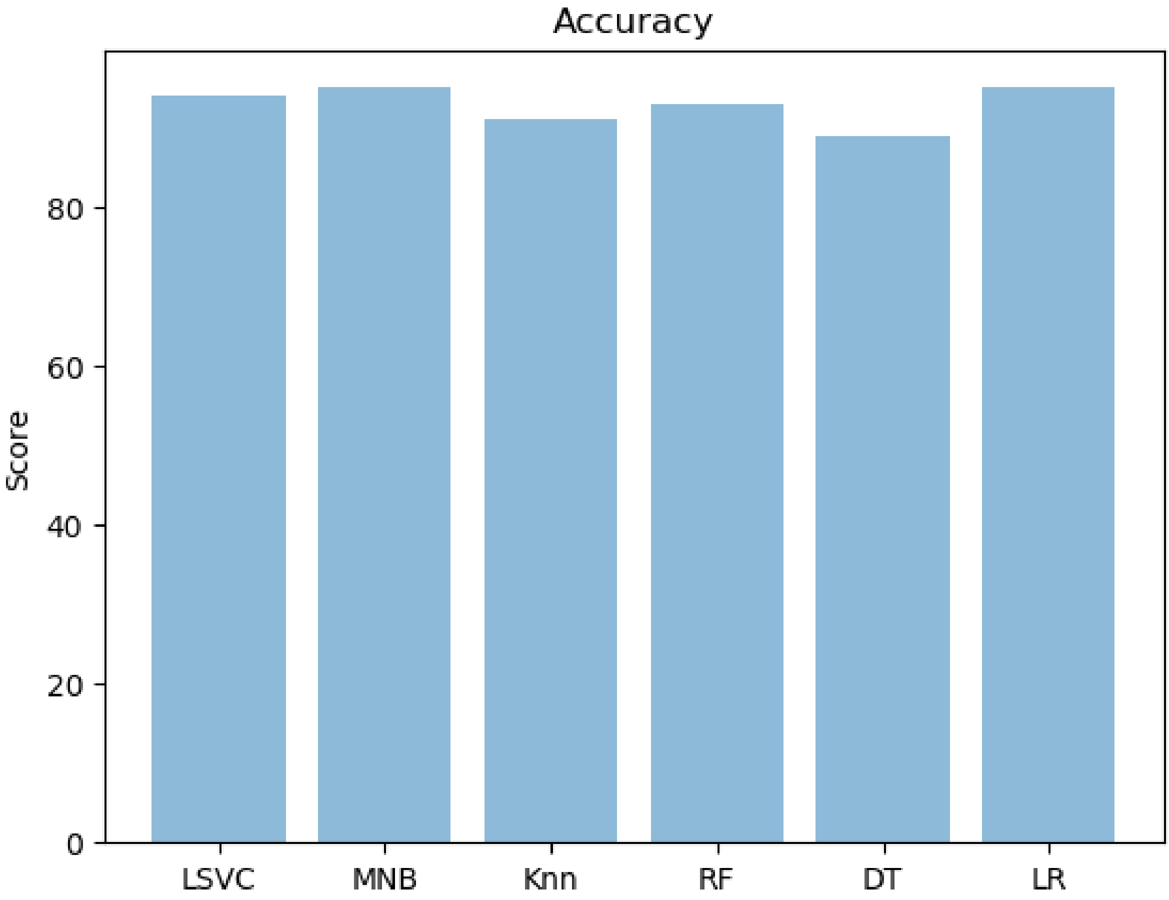


Figure 3.3.24: Histogram of Accuracy for 1500 data

3.5 Implementation requirements:

As we've completed every essential step to build this project Therefore it need some sorts of requirements which are crucial. These following requirements were used in our developed project.

Software or Hardware Requirements:
1. OS (Windows / Linux / Mac)
2. RAM (Minimum requirements of 4 GB)
3. HDD(Minimum requirements of 128 GB)

Developing Tools:
1. Jupyter Notebook / Google colab / Pycharm
2. Microsoft Word / Latex / Text Editor
3. Drawing Tools
4. Python environment.
5. Microsoft Excel.

CHAPTER 4: EXPERIMENTALRESULTSANDDISCUSSION

4.1 Experimental setup:

For experimenting we had to setup the python environment on our computer at first. Then we used a complier for coding where we chose Jupyter Notebook editor. Then we have used different python libraries including Numpy , Pandas, Matplotlib , Seaborn, NLTK, Cufflinks, scikit-learn, OS , Warnings, Strings, Wordcloud etc. For importing the above libraries, we had to install different packages. Sometimes we faced import error then we had to again re-install the requirements. This is how we implemented our setup.

4.2 Experimental Results &Analysis:

For checking initial accuracy we've implemented VADER at first but that was far away from our expected outcome. We were tensed to find out the appropriate classifiers which could be compatible with our dataset. After researching over 1 month we've chosen 6 supervised classifiers which are- Linear SVC, Multinomial Naive Bayes, Logistic Regression, Knn, Random Forest, Decision Tree where every single classifier performed well and most especially among all of them Linear SVC and Multinomial Naive bayes performed better than other classifiers. Comparing of six classifiers are given below:

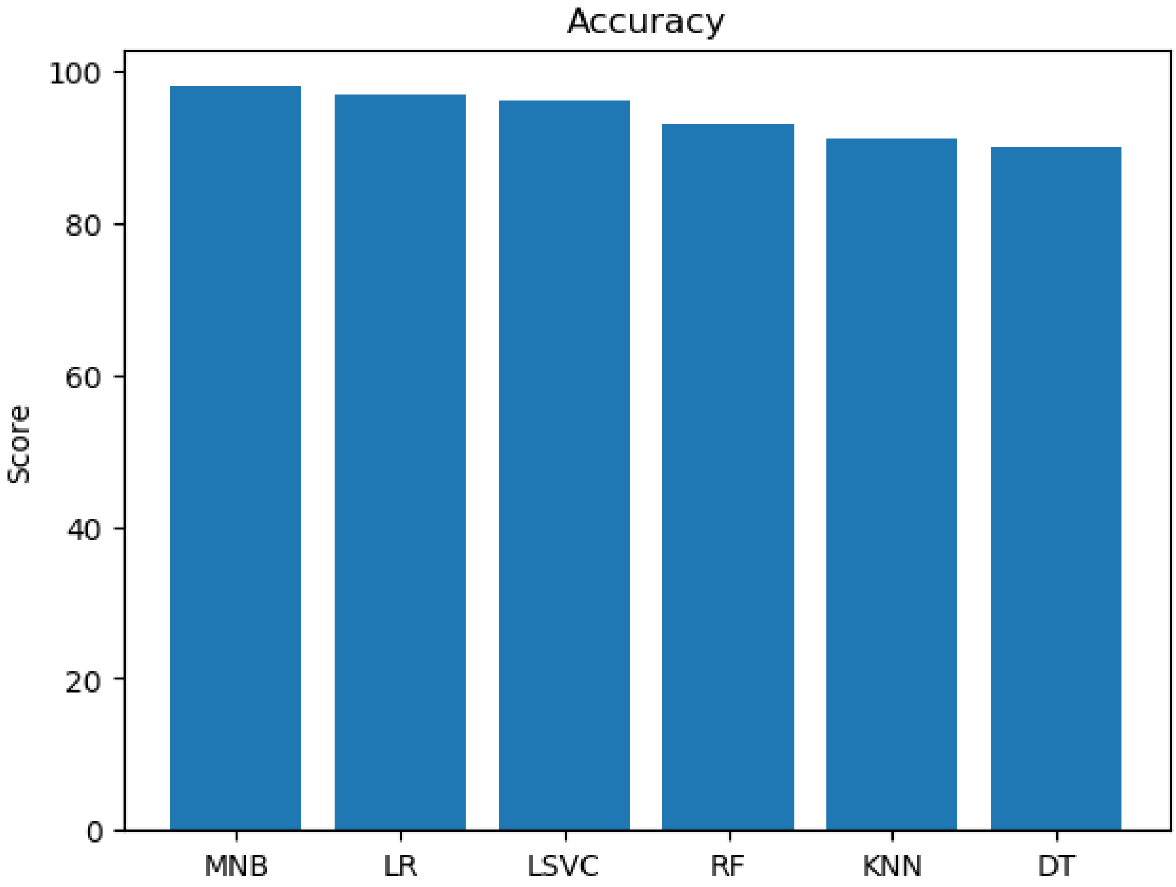


Figure 3.3.25: Histogram of Accuracy for 2000 data

4.3 Result Discussion: After increasing our dataset from 1500 data to 2000 we data our accuracy also increased. The highest accuracy we obtained from Naïve Bayes and Linear SVC which was 98% on our dataset. Related precision, recall, f1-score were very near to our accuracy. Therefore, we can say that our model performed very well on our dataset.

Precision: The number of correct documents returned by machine learning model is known as precision . It referred as the percentage of that classifier labeled as positive are actually positive. That means precision is the number of positive class prediction that are actually positive class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

-----------------------------------(2)

Recall: Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

-----------------------------------(3)

Here,
TP = Value of true positive
TN = true negative
FP = Value of false positive
FN = Value of false negative
F1-score: F1-Score refereed as Harmonic mean of precision and recall.

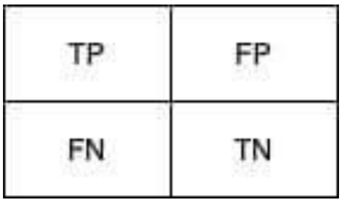
$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

-----------------------------------(4)

Accuracy: The number of correct predictions made as a ratio of all predictions made.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad \text{----------------------------------}(5)$$

Confusion Matrix:A confusion matrix is a technique for summarizing the performance of a classification algorithm.



TP	FP
FN	TN

Figure 3.3.26: Confusion Matrix

CHAPTER 5: IMPACTONSOCIETY, ENVIRONMENTANDSUSTAINABILITY

5.1 Impact on society:

Our aim is to remove depression from our society. Millions of people are suffering from depression all over the world. Prolong depression leads to suicide. Therefore the suicidal rate is increasing day by day. Due to covid-19 pandemic this situation is getting worst. From teenage to old person depression had spread hilariously. Our research work will help people to find out there depression level which will help them to improve their mental condition.

5.2 Ethical aspects:

Immensely use of technology has turned off our mental peace and also decreased our mental health. Nowadays people are spending their sheer amount of time in social media. They are showing to everyone that they are happy with their lives but the harsh reality is that most of them are not actually happy. Covid-19 pandemic has forced people to stay at their home this has also increased depression among the people. Our main goal is to find out the depression of each people and nurture them with proper treatment to increase their mental condition.

5.3 Sustainability plan: To make sustainable one of the effective ways can be "Early Detection". As we want to make a web version of our work. Therefore, on that web application whenever a user post or share some contents automatically a pop-up message will appear on screen which will indicate depression level. Second sustainable plan can be "Easy Accessibility". As most of people in our country is below the poverty line therefore that would be tough to purchase a paid apps therefore no alternative to make an open platform that can help people from all over the country. Our web application app can have a feature that provides free suggestions to the need of user and provide free mental health counseling. Third option can be "Lowered Fear of Stigma". In our surroundings some sorts of people who may not to discuss or share their problem publicly therefore our web application can have a Gender base selection system on which patient can choose their preferable selection. So, these can be sustainable plan of our work.

CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATIONANDIMPLICATIONFORFUTURE RESEARCH

6.1 Summary of the study:

There were too many research works had been done previously on this topic but most of the paper did not worked on real life data. We have collected our data from Facebook, twitter and instagram which had provided us more realistic results. First, we have found the polarity of each sentence by using Valence Aware Dictionary for Sentiment Reasoning (VADER). The accuracy we got was not satisfactory. Then we have cleaned and preprocessed our data by implementing different data preprocessed techniques such as removing of stop words, dropping of null values, removing punctuations, removing empty string, Lemmatization and steaming , concatenation of word to sentence ,count vectorize, tf-idf vectorizer and many more. Then we had to choose best classifiers which will perform better on our dataset. After researching a lot we had found six classifiers named as Multinomial Naïve Bayes, RF, Linear SVC,DT, Logistic regression and , KNN . Among all of them Multinomial Naïve Bayes and Linear SVC performed very well on our dataset.

6.2 conclusion:

Nowadays people are suffering from prolong depression due to covid-19 pandemic situation. It's getting worst day by day. Most of the people commit suicide due to prolong depression. They think that nobody can help them to get out from their hilarious and worst situation. In our work, we proposed a methodology which followed Supervised technique. We optimize the solution based on various classifiers. For increasing the correctness of our data we pre-processed very carefully on which we used top-notch advanced functionalities. Extraction feature was the hardest phase to us but we handled it optimally which helped us to increase accuracy level. The outcome that we've achieved is genuinely stimulate. Perhaps, the methodology will be an appropriate guideline in future of further in Depression analysis area.

6.3 Future work:

At present we are working on a small dataset which is consisting of 2000 sentences. In future we will work on a large dataset near around 10000 sentences. We will make a web and android app version of our work. We will make our application in such a way where a medical team will work with the depressed people to develop their condition. We have a plan to add live chatting, audio and video calling section where depressed people can chat with themselves and also with experts. This will reduce their loneliness and will help them to lead a normal life. Our aim is to free our society from depression and suicide.

Reference:

1. M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, 2017, pp. 858-862, doi: 10.1109/ISS1.2017.8389299.

2. P. Arora and P. Arora, "Mining Twitter Data for Depression Detection," 2019 International Conference on Signal Processing and Communication (ICSC), NOIDA, India, 2019, pp. 186-189, doi: 10.1109/ICSC45622.2019.8938353.

3. N. A. Asad, M. A. Mahmud Pranto, S. Afreen and M. M. Islam, "Depression Detection by Analyzing Social Media Posts of User," 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), Dhaka, Bangladesh, 2019, pp. 13-17, doi: 10.1109/SPICSCON48833.2019.9065101.

4. KantineeKatchapakirin, Asian Institute of Technology (AIT) , Academia Edu, <https://www.academia.edu/37672215/Facebook_Social_Media_for_Depression_Detection_in_the_Thai_Community>

5. S. A. S. A. Kulasinghe, A. Jayasinghe, R. M. A. Rathnayaka, P. B. M. M. D. Karunarathne, P. D. Suranjini Silva and J. A. D. C. Anuradha Jayakodi, "AI Based Depression and Suicide Prevention System," 2019 International Conference on Advancements in Computing (ICAC), Malabe, Sri Lanka, 2019, pp. 73-78, doi: 10.1109/ICAC49085.2019.9103411.

6. R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter and A. K. Das, "An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 360-364, doi: 10.1109/CCOMS.2019.8821658.

7. N. Tabassum and M. I. Khan, "Design an Empirical Framework for Sentiment Analysis from Bangla Text using Machine Learning," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ECACE.2019.8679347.

8. N. J. Ria, S. A. Khushbu, M. A. Yousuf, A. K. M. Masum, S. Abujar and S. A. Hossain, "Toward an Enhanced Bengali Text Classification Using Saint and Common

Form," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225358.

9. S. Nigam, A. K. Das and R. Chandra, "Machine Learning Based Approach To Sentiment Analysis," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida (UP), India, 2018, pp. 157-161, doi: 10.1109/ICACCCN.2018.8748848.

10. M. R. Hasan, M. Maliha and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036670.

11. V. Goel, A. K. Gupta and N. Kumar, "Sentiment Analysis of Multilingual Twitter Data using Natural Language Processing," 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2018, pp. 208-212, doi: 10.1109/CSNT.2018.8820254.

12. E. Laoh, I. Surjandari and N. I. Prabaningtyas, "Enhancing Hospitality Sentiment Reviews Analysis Performance using SVM N-Grams Method," 2019 16th International Conference on Service Systems and Service Management (ICSSSM), Shenzhen, China, 2019, pp. 1-5, doi: 10.1109/ICSSSM.2019.8887662.

13. Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017). Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. 2017 International Conference on Information and Communication Technology Convergence (ICTC). doi:10.1109/ictc.2017.8190959

14. D. Mahajan and D. Kumar Chaudhary, "Sentiment Analysis Using Rnn and Google Translator," 2018 8th International Conference on Cloud Computing, Data Science &

Engineering (Confluence), Noida, 2018, pp. 798-802, doi: 10.1109/CONFLUENCE.2018.8442924.

15. D. Tanna, M. Dudhane, A. Sardar, K. Deshpande and N. Deshmukh, "Sentiment Analysis on Social Media for Emotion Classification," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 911-915, doi: 10.1109/ICICCS48265.2020.9121057.

16. Wajdi Zaghouani, “A Large-Scale Social Media Corpus for the Detection of Youth Depression (Project Note),” Procedia Computer Science, Vol. 142, pp. 347-351, ISSN 1877-0509, 2018.

17. Sharma M., Pant B., Singh V., Kumar S., “STP:Suicidal Tendency Prediction Among the Youth Using Social Network Data,” Advances in Intelligent Systems and Computing, vol. 1162. Springer, 2021.

18. S. Nigam, A. K. Das, and R. Chandra, “Machine Learning Based Approach to Sentiment Analysis,” International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida (UP), pp. 157-161, 2018.

19. E. Laoh, I. Surjandari, and N. I. Prabaningtyas, “Enhancing Hospitality Sentiment Reviews Analysis Performance using SVM N-Grams Method,” 16th International Conference on Service Systems and Service Management (ICSSSM), pp. 1-5, 2019.

20. Hassan, A. U., et al., “Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression,” International Conference on Information and Communication Technology Convergence (ICTC), 2017.

21. D. Tanna, et al., “Sentiment Analysis on Social Media for Emotion Classification,” 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 911-915, 2020.

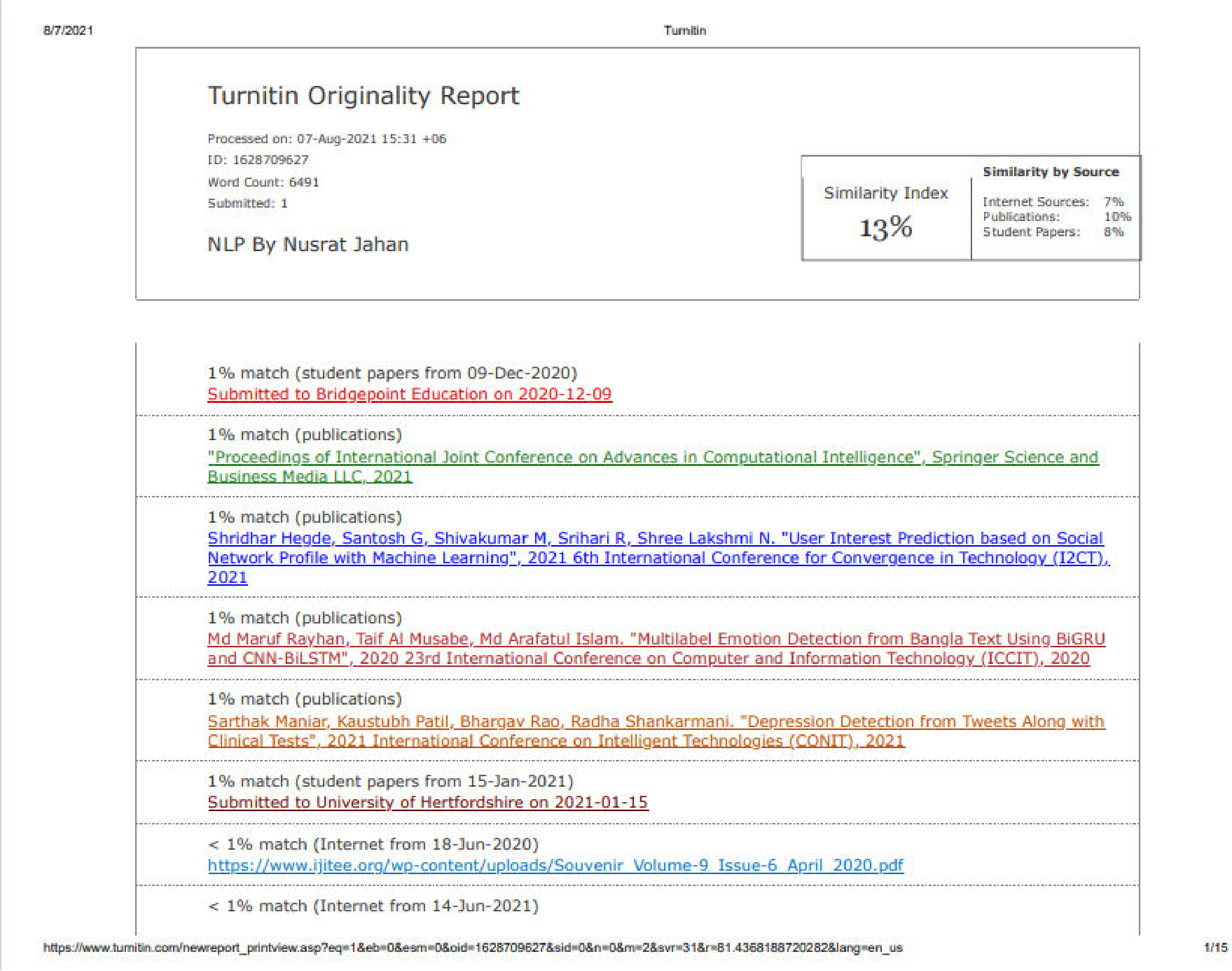
22. Hrishabh Patidar, and Jayesh Umre, “PREDICTING DEPRESSION LEVEL USING SOCIAL MEDIA POSTS” International Journal of Research –GRANTHAALAYAH, Vol 8(12), 234 – 237, December 2020.

23. M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of Depression-Related Posts in Reddit Social Media Forum,” IEEE Access, vol. 7, pp. 44883-44893, 2019.

24. Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi G. S., “A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis”. PLoS ONE 16(2): e0245909, 2021.

25. T. Basu and C. Murthy, “A feature selection method for improved document classication,"” in Advanced Data Mining and Applications. New York, Springer, pp. 296-305, 2012.

26. W. Xu-hui, S. Ping, C. Li and W. Ye, “A ROC Curve Method for Performance Evaluation of Support Vector Machine with Optimization Strategy,” 2009 International Forum on Computer Science-Technology and Applications, pp. 117-120, 2009.



8/7/2021 Turnitin

Turnitin Originality Report

Processed on: 07-Aug-2021 15:31 +06
ID: 1628709627
Word Count: 6491
Submitted: 1

NLP By Nusrat Jahan

Similarity Index	Similarity by Source	
13%	Internet Sources:	7%
	Publications:	10%
	Student Papers:	8%

1% match (student papers from 09-Dec-2020)
Submitted to Bridgepoint Education on 2020-12-09

1% match (publications)
"Proceedings of International Joint Conference on Advances in Computational Intelligence", Springer Science and Business Media LLC, 2021

1% match (publications)
Shridhar Hegde, Santosh G, Shivakumar M, Srihari R, Shree Lakshmi N. "User Interest Prediction based on Social Network Profile with Machine Learning", 2021 6th International Conference for Convergence in Technology (I2CT), 2021

1% match (publications)
Md Maruf Rayhan, Taif Al Musabe, Md Arafatul Islam. "Multilabel Emotion Detection from Bangla Text Using BiGRU and CNN-BiLSTM", 2020 23rd International Conference on Computer and Information Technology (ICCIT), 2020

1% match (publications)
Sarthak Maniar, Kaustubh Patil, Bhargav Rao, Radha Shankarmani. "Depression Detection from Tweets Along with Clinical Tests", 2021 International Conference on Intelligent Technologies (CONIT), 2021

1% match (student papers from 15-Jan-2021)
Submitted to University of Hertfordshire on 2021-01-15

< 1% match (Internet from 18-Jun-2020)
https://www.ijitee.org/wp-content/uploads/Souvenir_Volume-9_Issue-6_April_2020.pdf

< 1% match (Internet from 14-Jun-2021)

https://www.turnitin.com/newreport_printview.asp?eq=1&eb=0&esm=0&oid=1628709627&sid=0&n=0&m=2&svr=31&r=81.4368188720282&lang=en_us 1/15