DEPRESSION SENTIMENT ANALYSIS

A Project Work Synopsis

Submitted in the partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN
AIT-AIML

Submitted By:

Gauri Prabhakar

18BCS6201

Pranav Goel

18BCS6172

Under the Supervision of:

Mrs. Bhanu Priyanka

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413, PUNJAB

2021

Table of Contents

DECLARATION

We, **Gauri Prabhakar** and **Pranav Goel** students of '**Bachelor of Engineering in Computer Science and Engineering**', **session 2018-2022**, Apex Institute of Technology, Chandigarh University, Punjab, hereby declare that the work presented this Project Work titled '**Depression Sentiment Analysis**' is the outcome of our bona fide work and is correct to the best of our knowledge and this work have been undertaken taking care of Engineering Ethics.

It contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Gauri Prabhakar

(18BCS6201)

Pranav Goel

(18BCS6172)

Date: April 27th,2021

ABSTRACT

This synopsis briefly explains our project ‘DEPRESSION SENTIMENT ANALYSIS’ using machine learning. The project is at its base level but serves our main objective which is to make a revolution in Depression Sentiment Analysis. Existing systems use Convolutional neural network, Bayes theorem, and many other models to increase the accuracy of a tweet being treated as positive or negative.

The project aims at solving this problem by taking into account five models which are: Naive Bayes, Decision tree, Support vector machine, Kneighborsclassifier, Random Forest, and comparing their accuracy.
The goal is to feed a processed dataset that is well labeled and run a series of functions to identify the tweet.

The basic problem is the accuracy of these models is good but not at par and fully reliable.

The project is also getting its way and help from our respected professor Mrs. Bhanu. This synopsis will give you an overall idea of our project and the proposed model.

List of Figures

1. INTRODUCTION

'Depression is a constant feeling of sadness and loss of interest, which stops you from doing your normal activities. Different types of depression exist, with symptoms ranging from relatively minor to severe. Generally, depression does not result from a single event, but a mix of events and factors.' Technology especially social media has become one of the root causes of depression as it triggers an array of negative and overwhelming emotions. A study at the University of Pennsylvania suggests that the amount of time spent by people away from social media is inversely proportional to them developing symptoms of depression.
The project is an initiative towards identifying early depression in people based on the tweets they post in their day-to-day lives.

1.1 Problem Definition

The problem roots from the widespread negativity online on social media. The aim is to identify early signs of depression and then suggest the user, get help.

1.2 Project Overview/Specifications

The following section focuses on the requirements of the project.

1.3 Hardware Specifications

The hardware specifications are:

- Modern Operating System:
  - Windows 7 or 10
  - Mac OS X 10.11 or higher, 64-bit
  - Linux: RHEL 6/7, 64-bit
- x86 64-bit CPU (Intel / AMD architecture)
- 4 GB RAM
- 5 GB free disk space

1.4 Software Specifications

The software specifications are:

- Anaconda
- Jupyter Notebook
- Google Chrome
- Text viewer
- Spreadsheet
- Twitter developer account
- Python 3.6.1 or Higher
- A bunch of modules

2. LITERATURE REVIEW

Existing systems use Convolutional neural network, Bayes theorem, and many other models to increase the accuracy of a tweet being treated as positive or negative.
The basic problem is the accuracy of these models is good but not at par and fully reliable.

The project aims at solving this problem by taking into account five models and comparing their accuracy. The goal is to feed a processed dataset that is well labeled and run a series of functions to identify the tweet.

Another aspect of the project is to classify a tweet as neutral based on the labeled dataset ad keywords as -1,0,1.

2.1 Existing System

Existing systems use Convolutional neural network, Bayes theorem, and many other models to increase the accuracy of a tweet being treated as positive or negative.
The basic problem is the accuracy of these models is good but not at par and fully reliable.

The data fed to the model is highly flawed as it is very biased data, containing the word ‘depression’ or contains too many repetitive tweets.

The accuracy of the existing models is low and not at par with reality.
The existing state of the art uses way too simplistic models such as ‘logistic regression’ and a very simple dataset that is not up to date.

2.2 Proposed System

The project aims at solving this problem by taking into account five models which are: Naive Bayes, Decision tree, Support vector machine, Kneighborsclassifier, Random Forest, and comparing their accuracy.
The goal is to feed a processed dataset that is well labeled and run a series of functions to identify the tweet.

The model first downloads an updated Twitter API and then the data is retrieved and pre-processed.

Another aspect of the project is to classify a tweet as neutral based on the labeled dataset ad keywords as -1,0,1.

The project uses various modules and dataset which are available in python, including:
Nltk, csv, string, re, time, pandas etc.

All of the existing methods are correct and are great algorithms but they are being used independently. The algorithms that are being used are great but we tend to create a system that uses all of the main algorithms to recognize the best.

3. PROBLEM FORMULATION

In the existing systems, we can say that there is one common thing that all of them are majorly based upon one algorithm although the most efficient one but still not perfect.

As we all know that the decisions are more efficient when taken by a hundred minds rather than one. Similarly, there may be algorithms which can make the identification correct but also many other can make it wrong and we are not sure of which one is correct and which one is wrong. So it would be more efficient to make the decisions out of different algorithms and make the decision out of their mean results.

This could not only make it more efficient but also a dynamic method for identification out of every aspect and covering up the limitations of each other.

4. RESEARCH OBJECTIVE

The proposed research is aimed to carry out work leading to the development of an approach for Depression Sentiment Analysis.

The research objective is to find the algorithms which are the most efficient one can cover up each other's limitations which can help each other to achieve an accuracy that is more towards perfection.

The objective is to produce an updated dataset that is labeled as Positive that is 1, Negative that is -1, and neutral which is 0.

Advantages:

- Output generated based on updated Twitter API.
- Use of various up-to-date modules in python.
- Relatively the accuracy is at par.

Comparison is done based on 5 important models:

Naive Bayes.
Decision tree.
Support vector machine.
Kneighborsclassifier.
Random Forest.

Limitations:

- The computation time is large.
- Language is still a subjective matter, what might be positive to one may be negative to the other.

5. METHODOLOGY

The Flowchart:

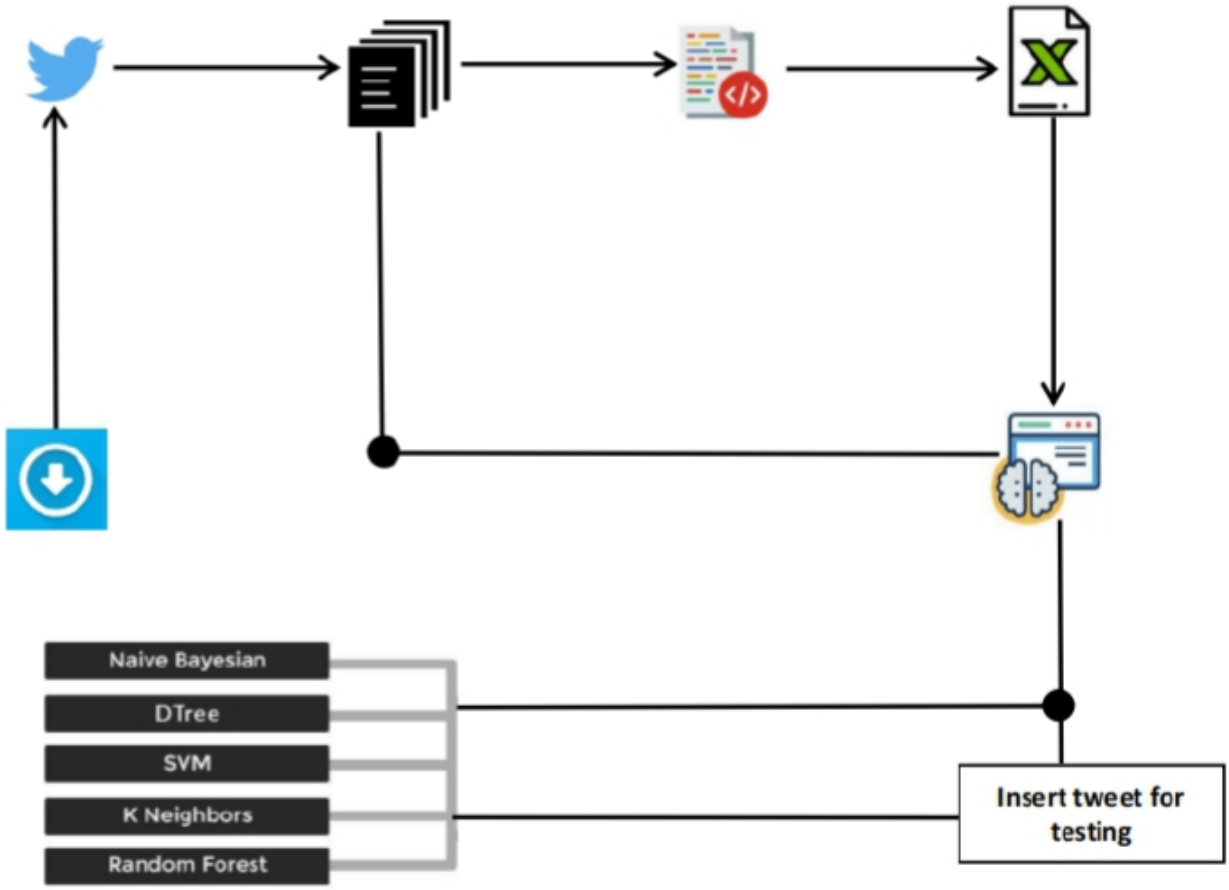

Figure 5.1 Methodology

The following methodology will be followed to achieve the objectives defined for the proposed research work:

Detailed study of datasets and algorithms will be done.

Installation and hands-on experience on existing approaches to depression sentiment analysis will be done. Relative pros and cons will be identified.

Various parameters such as depression, accuracy, stopwords will be identified to evaluate the proposed system.

First create a Twitter development account and extract consumer_key = '', consumer_secret = '', access_token = '', access_secret = ''.

Download the current tweets with the help of certain keywords such as 'depression', 'hopelessness', 'anxiety', 'stress', etc.

Then process the data to better fit the model.

Then pre-process the data from the dictionary and data set, the dictionary contains the polarity of the words, each word is tokenized and given a label, -1,0,1.

Each tweet consists of the summation of all the polarities of each word and is divided by the number of words in that particular tweet.

It shall look like this after extracting the data and processing it:

```
    print("\n*****************")
    print(predictt)
    print("******************")

runall()

#print("Input your tweet : ")
#inputtweet = input()
#
#datreeINPUT(inputtweet)
```

```
Naive Bayes  Accuracy :
 89.17885768634126 %
 Completion Speed 5.10621

Decision tree Accuracy :
 97.10166704910299 %
 Completion Speed 153.45589

Support vector machine Accuracy :
 91.88833735365208 %
 Completion Speed 7276.36305

Kneighborsclassifier Accuracy :
 82.28438353422621 %
 Completion Speed 587.23266

Random Forest Accuracy :
 48.87584020894772 %
 Completion Speed 9.67998
```

Figure 5.2 Accuracies

After pre-processing the file is saved in the directory 'processed_data/output.xlsx'.
This file contains: ID (tweet) and the Sentiment of each tweet is separated into 2 columns.

The Twitter dataset along with the sentiments is now ready.

Now for training and Predicting. The code will run through the output.xlsx file and at the same time recover the tweet corresponding to the id of each sentiment. using this we use the original data and feed them to our classifiers. When everything is done you should have all the AUC of each classifier listed in the console.

We can now even type a sample tweet like this:

```
Input your tweet :
I do not feel good today.

*****************
Neutral
*****************
```

Figure 5.3 Neutral Sentiment

Positive means the person is unlikely to have depression, negative means, the person is likely to have depression, and neutral means the person is in a gray state.

6. RESULTS

- **The Proposed model yields the accuracy and completion time as follows:**

```
Naive Bayes  Accuracy :
 89.17885768634126 %
 Completion Speed 5.10621

Decision tree Accuracy :
 97.10166704910299 %
 Completion Speed 153.45589

Support vector machine Accuracy :
 91.88833735365208 %
 Completion Speed 7276.36305

Kneighborsclassifier Accuracy :
 82.28438353422621 %
 Completion Speed 587.23266

Random Forest Accuracy :
 48.87584020894772 %
 Completion Speed 9.67998
```

Figure 6.1 Resultant Accuracies

- The model, predicts the sentiment correctly and overcomes the limitation of the existing state of the art.

- The model pre-processes the data efficiently.

```
    print("\n*****************")
    print(predictt)
    print("*****************")

runall()

#print("Input your tweet : ")
#inputtweet = input()
#
#datreeINPUT(inputtweet)
```

```
Naive Bayes  Accuracy :
 89.17885768634126 %
 Completion Speed 5.10621

Decision tree Accuracy :
 97.10166704910299 %
 Completion Speed 153.45589

Support vector machine Accuracy :
 91.88833735365208 %
 Completion Speed 7276.36305

Kneighborsclassifier Accuracy :
 82.28438353422621 %
 Completion Speed 587.23266

Random Forest Accuracy :
 48.87584020894772 %
 Completion Speed 9.67998
```

Figure 6.2 Proposed Algorithm's Speed

- The model labels data and specifies the sentiment of the tweet efficiently.

```
Input your tweet :
I do not feel good today.

*****************
Neutral
*****************
```

Figure 6.3 Resultant Sentiment

- **The Confusion Matrices:**

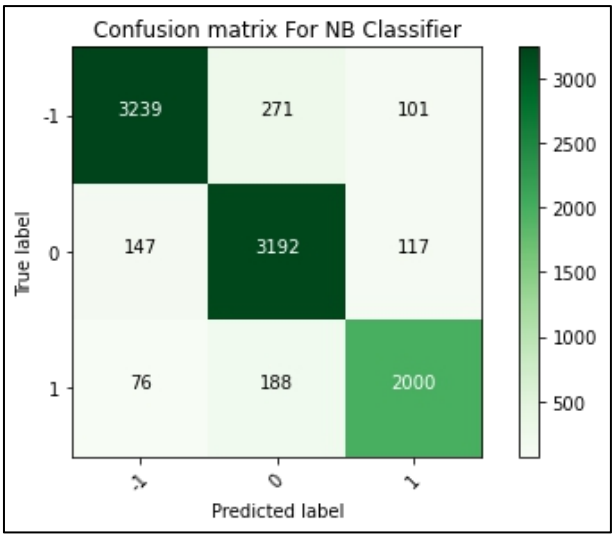

Figure 6.4 Confusion Matrix for NB Classifier

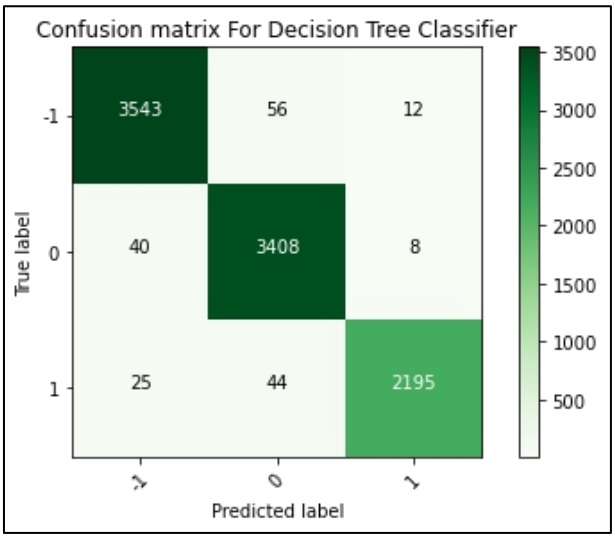

Figure 6.5 Confusion Matrix for Decision Tree Classifier

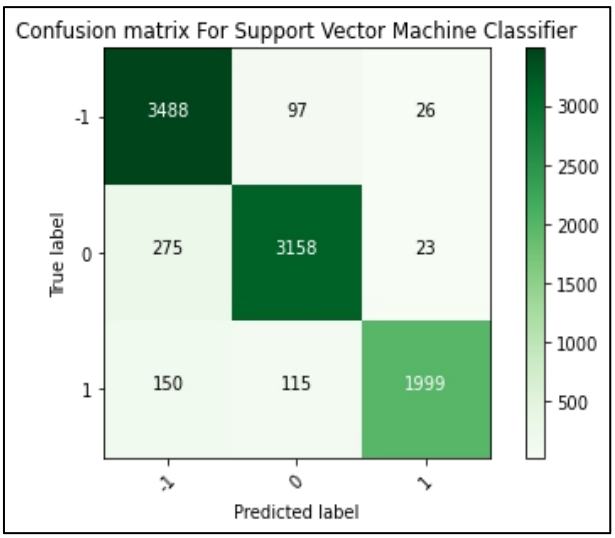

Figure 6.6 Confusion Matrix for SVM Classifier

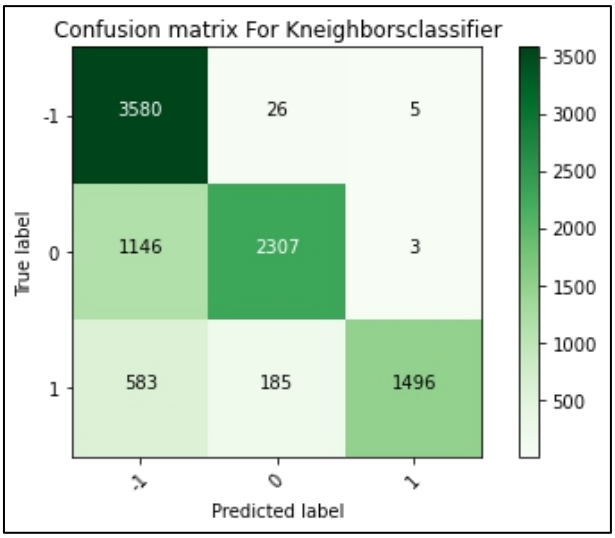

Figure 6.7 Confusion Matrix for K-neighbors Classifier

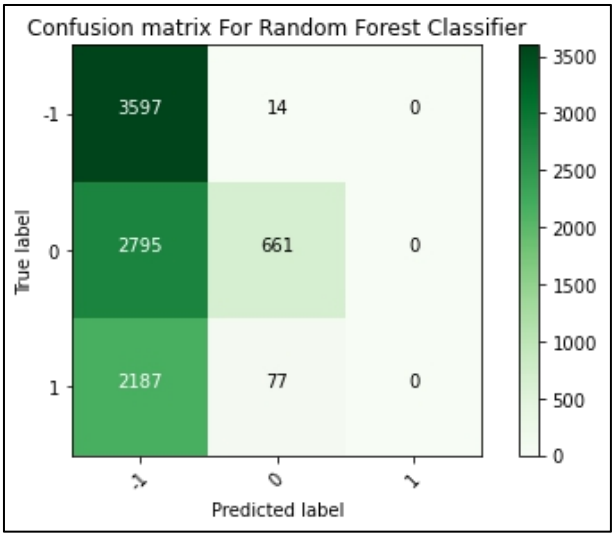

Figure 6.8 Confusion Matrix for Random Forest

7. CONCLUSION AND FUTURE SCOPE

Though the proposed model yields great accuracy as a result, the computation time is sacrificed as it is very large.

The future scope of the project:

The model can successfully be incorporated into the existing state-of-the-art products such as GBoard on Android, the Google Keyboard, which use federated learning to improve the experience of the user based on what they search.

If the user shows signs of depression the system can suggest the user use a self-care chatbot without hampering the user's privacy. The suggestion of a self-care bot can be an automatic feature that is integrated into GBoard upon depression detection and does not require the revelation of the user's identity.

We also aspire to produce a dataset that is specifically designed for depression identification based on the tweets. Since such data is not readily available and proves to be a major stumbling block/obstacle in the development of the project.

The flowchart:

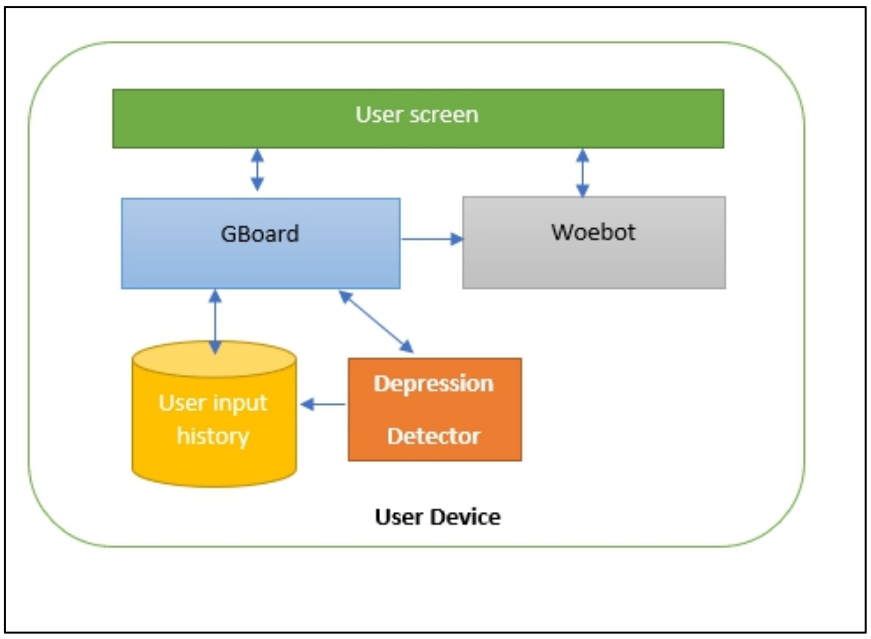

Figure 7.1 Future Scope

8. REFERENCES

- https://github.com/AshwanthRamji/Depression-Sentiment-Analysis-with-Twitter-Data
- https://arxiv.org/pdf/1607.07384.pdf
- https://www.researchgate.net/publication/318136574_Extracting_Depression_Symptoms_from_Social_Networks_and_Web_Blogs_via_Text_Mining
- https://vgpena.github.io/classifying-tweets-with-keras-and-tensorflow/
- Anne Bonner's Medium article You Are What You Tweet.
- Sentiment Analysis — TorchText