***A Project Report***

*on*

**DETECTION OF DEPRESSION ON TWITTER USING MACHINE LEARNING TECHNIQUES**

*carried out as part of the* ***Minor Project IT3270*** *Submitted*

by

**Vanshika Goyal**

**199302049**

**&**

**Riyanshi Bohra**

**199302110**

*in partial fulfilment for the award of the degree* *of*

**Bachelor of Technology**

in

**Information Technology**

Under the Guidance of

**Ms. Smaranika Mohapatra**



**School of Computing and Information Technology**

**Department of Information Technology**

**MANIPAL UNIVERSITY JAIPUR**

**JAIPUR-303007**

**RAJASTHAN, INDIA**

**May 2022**

**CERTIFICATE**

Date:

This is to certify that the minor project titled **DETECTION OF DEPRESSION ON TWITTER USING MACHINE LEARNING TECHNIQUES** is a record of the bonafide work done by **VANSHIKA GOYAL** (199302049) and **RIYANSHI BOHRA** (199302110) submitted in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Information Technologyof Manipal University Jaipur, during the academic year 2021-22.

**Guide Name**

*Ms. Smaranika Mohapatra, Department of Information Technology*

*Assistant Professor*

*Manipal University Jaipur*

**Dr. Pankaj Vyas**

*HoD, Department of Information Technology*

*Manipal University Jaipur*

**ABSTRACT**

The use of Social Media Sites like Twitter is expanding extremely fast. Nowadays, people tend to rely on these social media applications to share their emotions and feelings. Thus, this readily available content has become helpful for us to analyse the mental health of such users. We can apply various machine learning techniques on social media data to extract the mental health status of a user focusing on Depression.

Detecting texts that express negativity in the data is one of the best ways to detect depression. The Ensemble Learning approach for solving this problem has been enlightened. We aim to find and implement the most appropriate approach and algorithm to solve this problem.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No** | **Figure Title** | **Page No** |
|  |  |  |
| 3.1 | Flowchart Diagram | 5 |
| 3.3.1 | Word Cloud for Negative Words | 8 |
| 3.3.2 | Word Cloud for Positive Words | 8 |
| 4.3.1 | Confusion Matrix for BNB Model | 12 |
| 4.3.2 | Confusion Matrix for LR Model | 12 |
| 4.3.3 | Confusion Matrix for SVC Model | 13 |
| 4.3.4 | Sample text classification | 13 |
| 4.4 | Timeline Chart | 14 |

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Title** | **Page No.** |
| **1.** | **Introduction** | 1 |
| 1.1. | Introduction | 2 |
| 1.2. | Problem Statement | 2 |
| 1.3. | Objectives | 3 |
| 1.4. | Scope of Project | 3 |
| **2.** | **Background Detail** | 4 |
| 2.1. | Conceptual Overview / Literature Review | 4 |
| **3.** | **System Design & Methodology** | 5 |
| 3.1. | System Architecture | 5 |
| 3.2. | Development Environment | 6 |
| 3.3. | Methodology: Algorithm/Procedures | 6 |
| **4.** | **Implementation and Result** | 10 |
| 4.1. | Modules/Classes of Implemented Project | 10 |
| 4.2. | Implementation Detail | 10 |
| 4.3. | Results and Discussion | 11 |
| 4.4 | Month wise plan of work | 14 |
| **5.** | **Conclusion and Future Plan** | 15 |
|  | **References** | 16 |

**1. Introduction**

***1.1 Introduction***

Depression is a common mental illness and a leading cause of disability worldwide, which may cause suicide. Depression is typically diagnosed using face-to-face clinical depression criteria. However, in the early phases of depression, 70% of individuals would not seek medical help, potentially advancing their disease.

Because of its widespread use, social media offers a wealth of opportunities to improve the data available to mental health clinicians and researchers, resulting in a better-informed and equipped mental health field. Furthermore, people are negatively affected by contagious negative emotions in social networks, which can lead to depression and other mental diseases. Mental illness is an established risk factor for suicide; about 80% of individuals who try or die by suicide have been diagnosed with a mental disease. Depression is the most common mental disorder, although it has gone undetected or untreated due to a lack of acknowledgement or denial. Early detection of major depression symptoms and treatment with timely intervention can help to prevent the emergence of major depression.

***1.2. Problem Statement***

“To detect depression on twitter by analyzing tweets and using Machine Learning classification algorithms to predict whether a tweet is depressive or not.”

People increasingly rely on social media for sharing emotions, and daily life activities thus helpful for detecting their mental health. We aim to apply Machine Learning Techniques on Social Data of a user like Twitter feeds for performing analysis focusing on depression detection.

***1.3. Objectives***

Detecting earlier depression can be a huge step to address the mental illness and offer support to the people suffering from this terrible mental illness.

* Data analysis of tweets by Machine Learning Techniques:

Sentiment analysis is the process to identify the tone of the text is either positive, negative, or neutral. This can allow identifying the state of the user whether they are in a positive mood or negative mood. The sentiment analysis technique is applied to each tweet to identify the sentiment score and labeled them as positive, negative, or neutral. Sentiment analysis in Twitter data quantifies the status of a tweet or comment by the user by calculating the sentiment scores.

The model will be written in python and it will tell whether a given tweet is depressive or not.

* Detection of depressive tweets:

After analyzing the data and finding the sentiment score on tweets via sentiment analysis technique, the machine learning algorithm enables to classify the tweets whether it is depressive or not depressive based on the sentiment score labels on tweets. We implement a machine-learning algorithm to detect depression based on the tweets data which was labeled with sentiment scores.

***1.4. Scope of Project***

The areas of study for our project are machine learning and artificial intelligence using python. Using a machine learning approach to detect depression will surely help social media users for detecting and predicting depression risk. A machine learning approach like classification under supervised learning can analyse and build a model based on social media dataset. There are many factors like users' posts, tweets, replies, post time, emotions, etc. which can contribute to detecting depression. Before making a model, we will do exploratory data analysis including data preprocessing on our dataset to thoroughly understand it.

**2. Background Detail**

***2.1. Conceptual Overview / Literature Review***

In this project, we analyze the dataset by first preprocessing the data, and then creating Word clouds for Positive and Negative tweets and see which words occur the most. Then we split the data into a Training and Testing set. We train the data in different classifiers and check which classifier has the most accuracy.

**Classifiers used-**

Support Vector Machine(SVM), Bernoulli Naive Bayes (BernoulliNB), and Linear Regression (LR) are some of the widely used algorithms in machine learning classification tasks. Out of these, Linear Regression demonstrated the best performance.

1. Bernoulli Naive Bayes Classifier (BernoulliNB) :

BernoulliNB is based on the “Bayes’ theorem” in probability. As a requirement of this theorem, NB can be applied only if the features are independent of each other. This is used for discrete data and it works on Bernoulli distribution. The main feature of Bernoulli Naive Bayes is that it accepts features only as binary values like true or false, yes or no, success or failure, 0 or 1 and so on. So when the feature values are binary we know that we have to use Bernoulli Naive Bayes classifier.



2. Support Vector Machine (SVM)**:**

SVM is a supervised learning model that underlines two different classes in a high-dimensional space. The training algorithm in SVM creates a potential hyperplane, which divides the cases from the two classes. It escalates the distance between the divided hyperplane and the training examples closest to the hyperplane. SVM can provide the prediction and determine which side of the hyperplane an object inclines; as shown in Fig. 1. Although SVM can tolerate the data outlier, it is computationally inefficient and sensitive to the kernel hyperparameters.

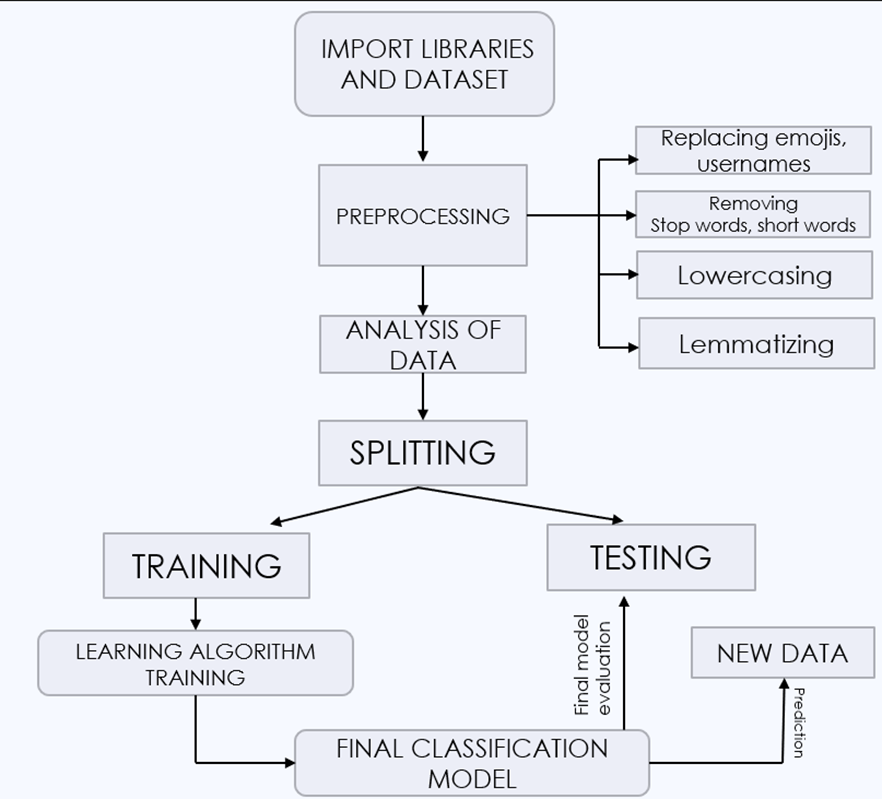
3. Logistic Regression (LR):

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

**3. System Design & Methodology: -**

***3.1. System Architecture***

Here is the flowchart for our Project-

****

**Fig. 3.1. Flowchart Diagram**

***3.2. Development Environment***

We will achieve our above-mentioned objectives by using a programming language called Python and by using Machine Learning techniques & algorithms. Following are the detailed software requirements for the development of our project:

· Kaggle(for the dataset)

· Jupyter Notebook which can be opened through Anaconda prompt.

· Wordcloud, a visualization tool.

Minimum System Requirements: -

· Processors: Intel Atom® processor or Intel® Core™ i3 processor

· Disk space: 1 GB

· Operating systems: Windows\* 7 or later, macOS, and Linux

· Python\* versions: 2.7.X, 3.6.X

***3.3. Methodology: Algorithm/Procedures***

The framework starts with data collection via Twitter scraper tools, which is then stored in a. csv file. The raw data will be cleaned and we start to do data pre-processing. The data will be normalized. The data will then be examined using sentiment analysis to generate a word score. The data are fed into different classifiers. The data is then split into the train and test sets. To ensure that the classifier learns, the training data is used for model construction. Once the model has learned about the data for evaluation, the test data will be fed into it. The accuracy results of the classifiers will be compared to determine which algorithm outperforms well.

The methodology of the project can be broadly classified into the following steps-

1) Data extraction and Data collection

Data Collection is the process of gathering and measuring information usually with software.

Data Extraction is where data is analyzed to retrieve relevant information from data sources (like a database) in a specific pattern. Further data processing is carried out, including adding metadata and data integration.

The dataset used for this project contains 1,600,000 tweets extracted using the **Twitter API**. The tweets have been labeled **(0 = negative, 4 = positive)** and can be used to determine sentiment.

The Dataset contains the following fields:

1. sentiment: the classification of the tweets *(0 = negative, 4 = positive)*

2. ids: tweet id

3. date: tweet date

4. flag: The query.

5. user: username of tweeps

6. text: the text of the tweet

For this project only the sentiment and text fields are needed, so we drop the rest.

Furthermore, the sentiment field will be updated to reflect the new sentiment format (0 = Negative, 1 = Positive).

2) Data Preprocessing

At this stage, the text is preprocessed and prepared for analysis and model development.

The Preprocessing steps taken are:

1. Lower Casing: Each text is converted to lowercase.

2. Replacing URLs: "http", "https" & "www" are replaced with "URL".

3. Replacing Emojis: Emojis are replaced with a pre-defined dictionary alternative

4. Replacing Usernames: Usernames are standardized.

5. Removing Non-Alphabets: Only retaining Digits, Alphabets and space.

6. Removing Consecutive letters: 3 or more consecutive letters are shortened

7. Removing Short Words: Words with length less than 2 are removed.

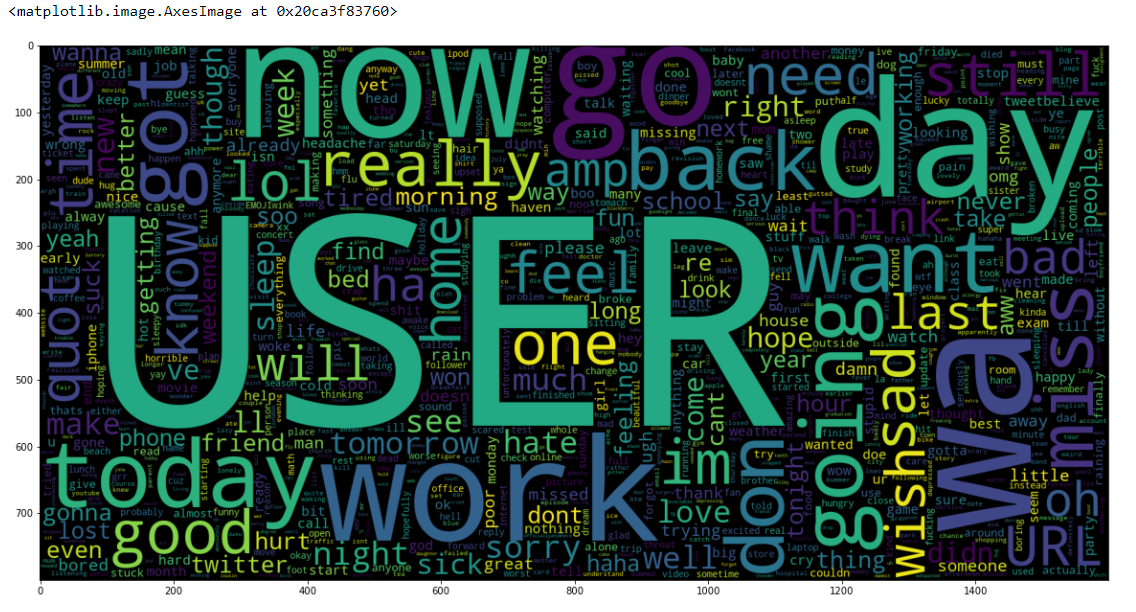
8. Removing Stopwords: Stopwords are the English words which does not add much meaning to a sentence.

9. Lemmatizing: Lemmatization is the process of converting a word to its base form.

3) Data Visualization

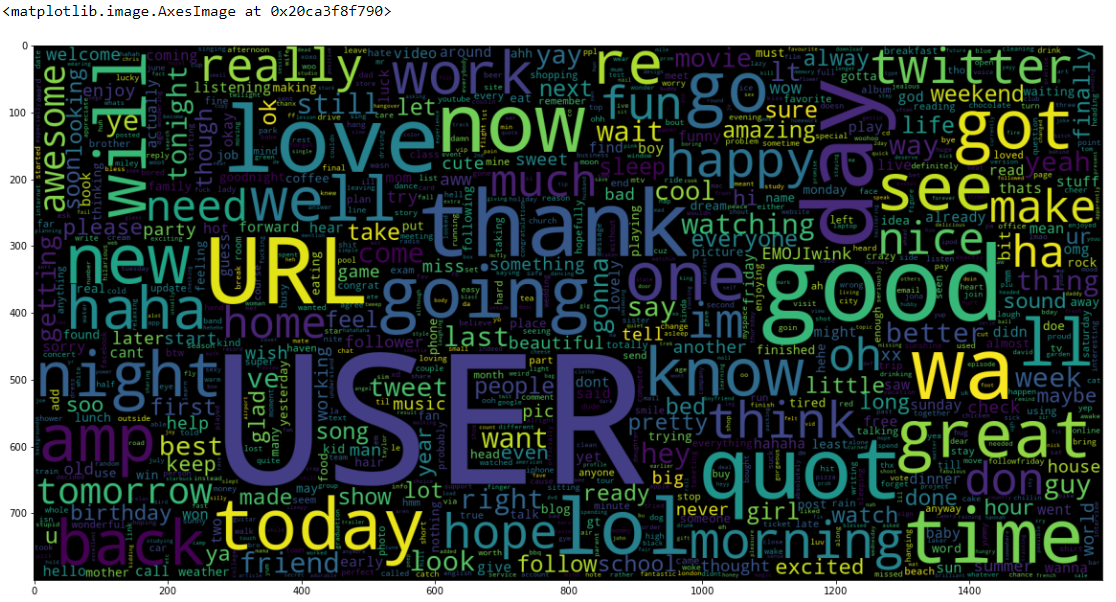
We'll use visualization to quickly explore the preprocessed data in order to find any hidden patterns, identify anomalies, and check assumptions. We'll use our dataset to create **Word Clouds** for **Positive and Negative** tweets to determine which words appear most frequently.

Word Cloud for Negative words-



**Figure 3.3.1**

Word Cloud for Positive words-



**Figure 3.3.2**

4) Data Analysis

In this project, we analyze the dataset by first preprocessing the data, and then creating Word clouds for Positive and Negative tweets and see which words occur the most. Then we split the data into a Training and Testing set. We train the data in different classifiers and check which classifier has the most accuracy.

5) Classification using Machine Learning Algorithms

The Classification algorithm is a Supervised Learning method that determines the class of new observations using training data. A program learns from a dataset or observations and then categorizes new observations into several classes or groups.

In this project, we used 3 machine learning models:

* Bernoulli Naive Bayes (BernoulliNB)
* Linear Support Vector Classification (LinearSVC)
* Logistic Regression (LR)

6) Training and Testing of Data

Train/Test is a technique for measuring your model's accuracy.

It's termed Train/Test because the data set is split into two parts: training and testing.

The first subset is known as the training data which is a portion of our actual dataset that is fed into the machine learning model to discover and learn patterns. Our model is so trained in this manner.

The testing data is the remaining subset. We can use it to evaluate the performance of our model.

The Preprocessed Data is split into;

* **Training set:** Containing 90% of the data.
* **Test set:** Containing 10% of the data

Precision-Recall Metrics is the evaluation metric employed in this research. We also used the Confusion Matrix to see how well our model performed across both classification types.

**4. Implementation and Result**

***4.1. Modules/Classes of Implemented Project***

The libraries used in our project are:-

* NumPy
* Pandas
* Wordcloud
* Seaborn
* Matplotlib
* Nltk
* Scikit-Learn

***4.2. Implementation Detail***

1. We first imported and extracted essential libraries.
2. Extraction of dataset from Kaggle
3. Next, we move on to data pre-processing for which we take the following steps:

* Lower Casing
* Removing Non-Alphabets, Consecutive letters, Short Words and Stop Words
* Replacing URLs, Emojis and Usernames

1. We analyze our data by using Word Cloud which visualizes the preprocessed data.
2. We then do Model Development and Evaluation on our dataset. We use three classifiers to create a model.
3. The evaluation metric used in this project is the Precision-Recall Metrics. We also plotted the **Confusion Matrix** to get an understanding of how our models are performing on both classification types.

For our Twitter sentiment analysis problem, we used three machine learning models in our project:

* Bernoulli Naive Bayes (BernoulliNB)
* Linear Support Vector Classification (LinearSVC)
* Logistic Regression (LR)

***4.3. Results and Discussions***

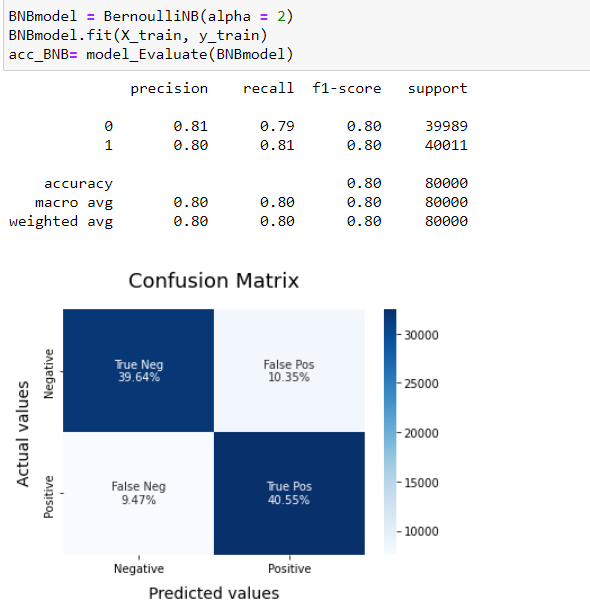
We plotted the **Confusion Matrix** to get an understanding of how our models are performing on both classification types.

After analyzing the Confusion Matrix, We can clearly see that the Logistic Regression Model performs the best out of all the different models that we tried. It achieves nearly 82% accuracy while classifying the sentiment of a tweet.

It should be noted, however, that the BernoulliNB Model is the quickest to train and predict. It also achieves 80% accuracy while classifying.

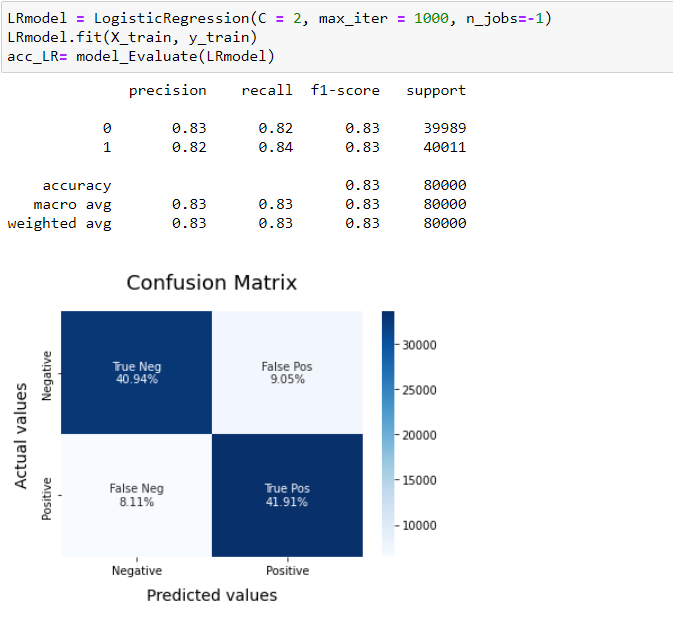
**Precision-Recall Metrics** is the evaluation metric used in this project.

For our Bernoulli Naive Bayes model,we get our confusion matrix as-



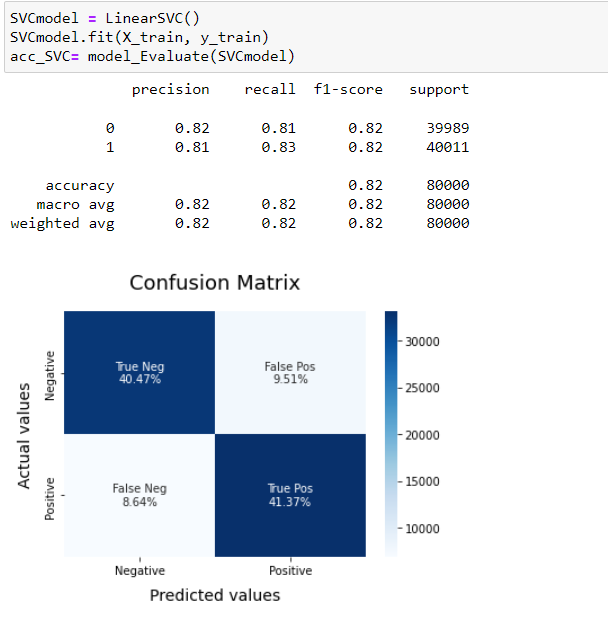
**Fig. 4.3.1 Confusion Matrix for BNB Model**

For Logistic Regression model,we get our confusion matrix as-

**

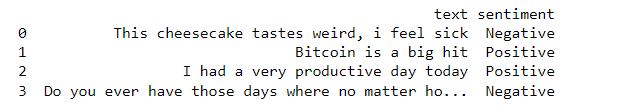
**Fig. 4.3.2 Confusion Matrix for LR Model**

For Linear Support Vector Classification Model, we get out Confusion Matrix as-



**Fig. 4.3.3 Confusion Matrix for SVC Model**

Sample Text we entered to classify-



**Figure 4.3.4 Sample text classification**

***4.4. Month wise plan of work (Progress Chart/Time Line Chart)***

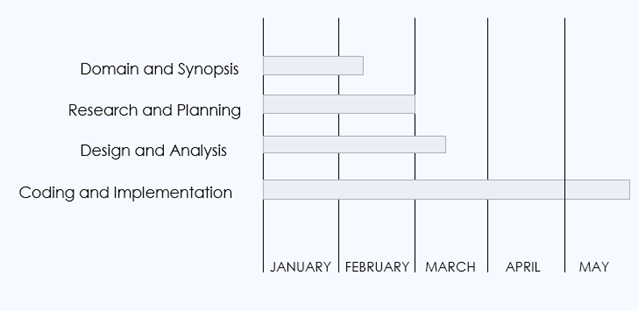


Fig. 4.4. Timeline Chart

**5.** **Conclusion and Future Plan: -**

This project defines a classification problem as identifying whether a person is depressed or not, based on their tweets.

This study used a dataset taken from twitter to explore various methods of early detection of depression based on machine learning. We performed a thorough analysis of the dataset to characterize the subjects’ behavior based on their tweets.

Different machine learning algorithms are used for the detection of depression. Many preprocessing steps are performed, including removal of stopwords and short words and lemmatizing. This project can be considered as a step toward building a complete social media-based platform for analyzing and predicting mental and psychological issues and recommending solutions for these users.

This study can be extended in the future by considering more ML models that might have much better accuracy for the classification problem and to find a more dependable way to measure the features’ impact.

**REFERENCES**

1. M. R. Islam, M. A. Kabir, A. Ahmed, A. R M. Kamal, H. Wang and A. Ulhaq, "Depression detection from social network data using machine learning techniques", vol. 6, pp. 1-12, 2018.
2. Holleran SE, “The early detection of depression from social networking sites.” Tucson: The University of Arizona, 2010.
3. Ang, L. (2011) Community relationship management and social media. Journal of Database Marketing and Customer Strategy Management, 31–38, doi:10.1057/dbm.2011.3.
4. De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E, “Predicting depression via social media”, ICWSM. 13, 1-10 (2013).
5. Harman, G., & Dredze, M. H. Measuring post traumatic stress disorder in Twitter. Proceedings of ICWSM, (2014).
6. Benton, A., Mitchell, M., & Hovy, D. Multi-task learning for mental health using social media text. Preprint at arXiv:1712.03538 (2017).
7. Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. Recognizing depression from twitter activity in Recognizing depression from twitter activity 3187-3196 (ACM, 2015).
8. Jamil, Z., Inkpen, D., Buddhitha, P., & White, K. Monitoring tweets for depression to detect at-risk users. CLPsych, 32, (2017).
9. Reece, A.G., et al. “Forecasting the onset and course of mental illness with Twitter data”, Sci. Rep. 7(1),13006 (2017).
10. Nadeem, M. Identifying depression on Twitter. Preprint at arXiv:1607.07384 (2016)
11. "How to Get Use & Benefit From Twitter's API", <https://blog.hubspot.com/website/how-to-use-twitter-api>.
12. <https://www.javatpoint.com/machine-learning>
13. Twitter Data Mining: Analyzing Big Data Using Python | Toptal, <https://www.toptalcom/python/twitter-data-mining-using-python>.
14. O’Dea B, et al. Detecting suicidality on Twitter. Internet Interv. 2015;2(2):183–8.
15. S. Dutta, D. Sarkar, S. Roy, D. K. Kole and P. Jana, "A Study on Herd Behavior Using Sentiment Analysis in Online Social Network," 2021 International Conference on Communication, Control and Information Sciences (ICCISc), 2021, pp. 1-6, doi: 10.1109/ICCISc52257.2021.9484918.