VISVESVARAYA TECHNOLOGICAL UNIVERSITY
BELAGAVI ,590018,Karnataka.

INTERNSHIP REPORT ON
Sentiment Analysis Of Lockdown In USA During Covid-19 A Case Study On Twitter using ML
Submitted inpartial fulfillment for the award of degree(18CSI85)

BACHELOR OF ENGINEERING IN
COMPUTER SCIENCE

Submitted by:

Shrusti S Ambudkar

USN: 2VD19CS050

Varcons Technologies Pvt Ltd

Conducted at
Varcons Technologies Pvt Ltd

KLS Vishwanathrao Deshpande Institute of Technology,Haliyal.
Computer Science and Engineering ,Uttara Kannada District, Haliyal.

CERTIFICATE

This is to certify that the Internship titled **"Sentiment Analysis of lockdown In USA During Covid-19 A case study on Twitter using ML"** carried out by **Miss** Shrusti S Ambudkar, bonafide student of **KLS Vishwanathrao Deshpande Institute of Technology**, inpartial fulfillment for the award of **Bachelor of Engineering,in Computer Science** under Visvesvaraya Technological University, Belagavi, during the year 2022-2023. It is certified that all corrections/suggestions indicated have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of Internship prescribed for the course Internship/Professional Practice (18CSI85).

Signature of Guide   **Signature of HOD**   **Signature of Principal**

External Viva:

Name of the Examiner   Signature with Date

1)_____

2)_____

DECLARATION

I.Shrusti S Ambudkar, final year student of Computer Science and Engineering, KLS Vishwanathrao Deshpande Institute of Technology, Haliyal 581329. declare that the Internship has been successfully completed, in **Varcons Technologies Pvt Ltd**. This report is submitted in partial fulfillment of the requirements for award of Bachelor Degree in Computer Science, during the academic year 2022-2023.

Date: 24-09-2022   :

Place : Mudhol

USN: 2VD19CS050

NAME: Shrusti S Ambudkar

OFFER LETTER

INTERNSHIP OFFER LETTER

Varcons Technologies Pvt Ltd

Date: **2ⁿᵈ September, 2022**

Name: **Shrusti Suresh Ambudkar**
USN: **2VD19CS050**

Dear Student,

We would like to congratulate you on being selected for the **Machine Learning With-Python(Research Based)** Internship position with **Varcons Technologies Pvt Ltd**, effective Start Date **2ⁿᵈ September, 2022**, All of us are excited about this opportunity provided to you!

This internship is viewed as being an educational opportunity for you, rather than a part-time job. As such, your internship will include training/orientation and focus primarily on learning and developing new skills and gaining a deeper understanding of concepts of **Machine Learning With Python(Research Based)** through hands-on application of the knowledge you learn while you train with the senior developers. You will be bound to follow the rules and regulations of the company during your internship duration.

Again, congratulations and we look forward to working with you!

Sincerely,

Spoorthi H C
Director
VARCONS TECHNOLOGIES PVT LTD
213, 2ˢᵗ Floor,
18 M G Road, Ulsoor,
Bangalore-560001

ACKNOWLEDGEMENT

This Internship is a result of accumulated guidance, direction and support of several important persons. We take this opportunity to express our gratitude to all who have helped us to complete the Internship.

We express our sincere thanks to our Principal Dr.V.A. Kulkarni, for providing us adequate facilities to undertake this Internship.

We would like to thank our Head of Dept Prof Poornima Raikar, Computer Science and Engineering for providing us an opportunity to carry out Internship and for her valuableguidance and support.

We express our deep and profound gratitude to our guide, Prajawal S Madhav for there keen interest and encouragement at every step in completing the Internship.

We would like to thank all the faculty members of our department for the support extended during the course of Internship.

We would like to thank the non-teaching members of our dept, for helping us during the Internship.

Last but not the least, we would like to thank our parents and friends without whose constant help, the completion of Internship would have not been possible.

Name: Shrusti S Ambudkar
USN: 2VD19CS053

ABSTRACT

COVID-19 originally known as Corona Virus Disease of 2019, has been declared as a pandemic by World Health Organization (WHO) on 11th March 2020. Unprecedented pressures have mounted on each country to make compelling requisites for controlling the population by assessing the cases and properly utilizing available resources. The rapid number of exponential cases globally has become the apprehension of panic, fear and anxiety among people. The mental and physical health of the global population is found to be directly proportional to this pandemic disease. The current situation has reported more than twenty four million people being tested positive worldwide as of 27th August, 2020. Therefore, it is the need of the hour to implement different measures to safeguard the countries by demystifying the pertinent facts and information. This paper aims to bring out the fact that tweets containing all handles related to COVID-19 and WHO have been unsuccessful in guiding people around this pandemic outbreak appositely. This study analyzes two types of tweets gathered during the pandemic times. In one case, around twenty three thousand most re-tweeted tweets within the time span from 1st Jan 2019 to 23rd March 2020 have been analyzed and observation says that the maximum number of the tweets portrays neutral or negative sentiments. On the other hand, a dataset containing 226,668 tweets collected within the time span between December 2019 and May 2020 have been analyzed which contrastingly show that there were a maximum number of positive and neutral tweets tweeted by netizens.

Table of Contents

CHAPTER 1

COMPANY PROFILE

COMPANY PROFILE

## A Brief History of Varcons Technologies

Varcons Technologies, was incorporated with a goal "To provide high quality and optimal Technological Solutions to business requirements of our clients". Every business is a different and has a unique business model and so are the technological requirements. They understand this and hence the solutions provided to these requirements are different as well. They focus on clients requirements and provide them with tailor made technological solutions. They also understand that Reach of their Product to its targeted market or the automation of the existing process into e-client and simple process are the key features that our clients desire from Technological Solution they are looking for and these are the features that we focus on while designing the solutions for their clients.

Sarvamoola Software Services. is a Technology Organization providing solutions for all web design and development, MYSQL, PYTHON Programming, HTML, CSS, ASP.NET and LINQ. Meeting the ever increasing automation requirements, Sarvamoola Software Services. specialize in ERP, Connectivity, SEO Services, Conference Management, effective web motion and tailor-made software products, designing solutions best suiting clients requirements.

Varcons Technologies, strive to be the front runner in creativity and innovation in software development through their well-researched expertise and establish it as an out of the box software development company in Bangalore, India. As of are development company, they translate this software development expertise into value for their customers through their professional solutions.

They understand that the best desired output can be achieved only by understanding the clients demand better. Varcons Technologies work with their clients and help them to define their exact solution requirement. Sometimes even they wonder that they have completely redefined their solution or new application requirement during the brain storming session, and here they position themselves as an IT solutions consulting group comprising of high caliber consultants.

They believe that Technology when used properly can help any business to scale and achieve new heights of success. It helps Improve its efficiency, profitability, reliability; to put it in one sentence " Technology helps you to Delight your Customers" and that is what we want to achieve.

CHAPTER -2

ABOUT THE COMPANY

1. ABOUT THE COMPANY

Varcons Technologies is a technology Organization providing solutions for all web design and development, MYSQL, PYTHON Programming, HTML, CSS, ASP.NET and LINQ. Meeting the ever increasing automation requirements, Varcons Technologies specialize in ERP, Connectivity, SEO Services, Conference Management, effective web promotion and tailor- made software products, designing solutions best suiting clients requirements. The organization where they have a right mix of professionals as a stakeholders to help us serve our clients with best of our capability and with at par industry standards. They have young, enthusiastic, passionate and creative Professionals to develop technological innovations in the field of Mobile technologies, Web applications as well as Business and Enterprise solution. Motto of our organization is to "Collaborate with our clients to provide them with best Technological solution hence creating Good Present and Better Future for our client which will bring a cascading a positive effect in their business shape as well". Providing a Complete suite of technical solutions is not just our tag line, it is Our Vision for Our Clients and for Us, We strive hard to achieve it.

Products of Varcons Technologies.

Android Apps

It is the process by which new applications are created for devices running the Androidoperating system. Applications are usually developed in Java (and/or Kotlin; or other such option) programming language using the Android software development kit (SDK), but other development environments are also available, some such as Kotlin support the exact same Android APIs(and byte code),while other such as Go have restricted API access.

The Android software development kit includes a comprehensive set of development tools. These include a debugger, libraries, a handset emulator based on QEMU, documentation,sample code, and zutorials. Currently supported development platforms include computers running Linux (any modern desktop Linux distribution), Mac OS X 10.5.8 or later, and Windows 7 or later. As of March 2015, the SDK is not available on Android itself, but software development is possible by using specialized Android applications.

search engine optimization. The term web design is normally used to describe the design process relating to the front-end (client side) design of a website including writing mark up. Web design partially overlaps web engineering in the broader scope of web development. Web designers are expected to have an awareness of usability and if their role involves creating mark up then they are also expected to be up to date with web accessibility guidelines. Web design partially overlaps web engineering in the broader scope of web development.

Departments and services offered

Varcons Technologies plays an essential role as an institute, the level of education, development of student's skills are based on their trainers. If you do not have a good mentor then you may lag in many things from others and that is why we at Varcons Technologies givesyou the facility of skilled employees so that you do not feel unsecured about the academics. Personality development and academic status are some of those things which lie on mentor's hands. If you are trained well then you can do well in your future and knowing its importanceof Varcons Technologies always tries to give you the best.

They have a great team of skilled mentors who are always ready to direct their trainees in thebestpossiblewaytheycanandtoensuretheskillsofmentorsweheldmanyskilldevelopment programs as well so that each and every mentor can develop their own skills with the demandsof the companies so that they can prepare a complete packaged trainee.

Services provided by Varcons Technologies.

- Core Java and Advanced Java

- Web services and development

- Dot Net Framework

- Python

- Selenium Testing

- Conference /Event Management Service

- Academic Project Guidance

- On the Job Training

- Software Training

CHAPTER-3

INTRODUCTION

1. <u>INTRODUCTION</u>

Introduction to ML

Definition of Machine Learning: Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term "Machine Learning" in 1959 while at IBM. He defined machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed ". However, there is no universally accepted definition for machine learning. Different authors define the term differently. We give below two more definitions.

Machine learning is programming computers to optimize a performance criterion using example data or past experience . We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training dataor past experience. The model may be predictive to make predictions in the future, or descriptiveto gain knowledge from data.

The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

Problem Statement

Goal: Understand the working of Sentiment analysis and Improve the accuracy

CHAPTER-4

SYSTEM ANALYSIS

SYSTEM ANALYSIS

1.Existing Problem

The whole world is dealing with the coronavirus pandemic right now. But another crisis has also manifested itself along with the virus, which is as detrimental as the virus itself. That is the large flow of information regarding the virus in the form of tweets, blogs, news, and too little analysis of this tremendous amount of information flowing in and out of the system every moment. This information could sometimes be fake, mixed news or just some opinion about the pandemic. The spreading of such incomplete, inaccurate news would be the causeof mass hysteria and fear among the public and hence the need of the hour is to address and better understand this information crisis regarding the pandemic. The government, companies, and many organizations form their decisions to cater to the needs of the people, and hence understanding the right sentiment and opinions of the people towards the pandemic plays a crucial role in policymaking and creating appropriate business models thatare of necessity to the people.

2.Proposed System

The exponential rise in the social media usage by the people in the last decade to express their opinions, perspectives and also as a source of news rather than the traditional news has laid emphasis on the usage of Deep Learning and Artificial Intelligence methods in gauging information from these sites to analyze and extract sentiments that act as valuable sources of insights for corporate companies and the government alike in making policies andappropriate business models to match the trend of the public. Hence, the main purpose of this project lies in the creation of a Public Sentiment Analysis Dashboard that depicts the sentiment trend among the Indian public towards the pandemic. Twitter is a microblogging site that has gained popularity in recent years, for the governments, organizations, and people alike, as a platform to make official announcements, expressing moods, opinions towards current, ongoing issues. Hence the tweets, that are the short posts that are made onthis platform, act as our data set for creating this dashboard. The application of Deep Learning methods to analyze these tweets, gauge sentiments from them, and then representing these results on a live interactive dashboard is the main aim of this project.

3.Objective of the System

Understand the working of Sentiment analysis and Improve the accuracy.

CHAPTER -5

REQUIREMENT   ANALYSIS

REQUIREMENT   ANALYSIS

Hardware and other Requirement Specification

- Linux Operating System/Windows
- Python Platform(Anaconda2,Spyder,Jupyter)
- Modern Web Browser
- Twitter API
- Google API
- NLTK package

Software Requirement Specification

- Operating System: Windows 10.
- Tools Used: Python version 3.7

CHAPTER- 6

System   Analysis

DESIGN & ANALYSIS

Training the Machine

Training the machine is similar to feeding the data to the algorithm to touch up the test data. The training sets are used to tune and fit the models. The test sets are untouched, as a model should not be judged based on unseen data. The training of the model includes cross-validation where we get a well-grounded approximate performance of the model using the training data. Tuning models are meant to specifically tune The hyper parameters like the number of trees in a random forest. We perform the entire cross-validation loop on each set of hyper parameter values. Finally, we will calculate a cross-validated score, for individual sets of hyper parameters. Then, we select the best hyper parameters. The idea behind the training of the model is that we some initial values with the dataset and then optimize the parameters which we want to in the model. This is kept on repetition until we get the optimal values. Thus, we take the predictions from the trained model on the inputs from the test dataset. Hence, it is divided in the ratio of 80:20 where 80% is for the training set and the rest 20% for a testing set of the data

Data Scoring

The process of applying a predictive model to a set of data is referred to as scoring the data. The technique used to process the dataset is the Random Forest Algorithm. Random forest involves an ensemble method, which is usually used, for classification and as well as regression. Based on the learning models, we achieve interesting results. The last module thus describes how the result of the model can help to predict the probability of a stock to rise and sink based on certain parameters. It also shows the vulnerabilities of a particular stock or entity. The user authentication system control is implemented to make sure that only the authorized entities are accessing the results.

CHAPTER-7

IMPLEMENTATION

6. IMPLEMENTATION

Sentiment Analysis of a Tweet With Naive Bayes

Millions of tweets are posted every second. It helps us know how the public is responding to a particular event. To get the sentiments of tweets, We can use the Naive Bayes classification algorithm, which is simply the application of Bayes rule.

Bayes Rule

Bayes rule is merely describing the probability of an event on prior knowledge of the occurrence of another event related to it.
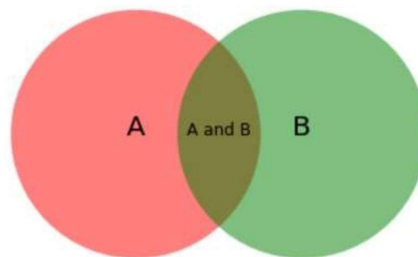
Then the probability of occurrence of event A given that event B has already occurred is

And for the probability of occurrence of event B given that event A has already occurred is

Using both these equations, we can rewrite them collectively as

$$P(B \mid A) = \frac{P(A \mid B) * P(B)}{P(A)}$$

A, B = Events

P(A|B) = Probability of 'A' given 'B' is True

P(A|B) = Probability of 'B' given 'A' is True

P(A), P(B) = The Independent Probabilities of 'A' and 'B'

Let's take a look at tweets and how we are going to extract features from them

We will be having two corpora of tweets, positive and negative tweets.

Positive tweets: 'I am happy because I am learning NLP,' 'I am happy, not sad.'

Negative tweets: 'I am sad, I am not learning NLP,' 'I am sad, not happy.'

Preprocessing

We need to preprocess our data so that we can save a lot of memory and reduce the computational process.

1. Lowercase: We will convert all the text to lower case. so, that the words like Learning and leaning can be taken as same words

2. Removing punctuations, URLs, names: We will remove the punctuations URLs and names or hashtags because they don't contribute to sentiment analysis of a tweet.

3. Removing stopwords: The stopwords like 'the', 'is' don't contribute in sentiment. Therefore these words have to be removed.

4. Stemming: The words like 'took', 'taking' are treated as the same words and are converted to there base words, here it is 'take'. This saves a lot of memory and time.

Probabilistic approach:

In order to get the probability stats for the words, we will be creating a dictionary of these words and counting the occurrence of each word in positive and negative tweets.

Word Count

```
+---------+-----+-----+
|  Word   | Pos | Neg |
+=========+=====+=====+
| i       | 3   | 3   |
| am      | 3   | 3   |
| happy   | 2   | 1   |
| because | 1   | 0   |
| learn   | 1   | 1   |
| nlp     | 1   | 1   |
| sad     | 1   | 2   |
| not     | 1   | 2   |
+---------+-----+-----+
| Nclass  | 13  | 12  |
+---------+-----+-----+
```

Let's see how these word counts are helpful in finding the probability of the word for both classes. Here the word 'i' occurred three times, and the total unique words in the positive corpus are 13. Therefore, the probability of occurrence of the word 'i' given that the tweet is positive will be

$$P(I/pos) = \frac{3}{13} = 0.24$$

$$P(w_i|class) = \frac{freq(w_i, class)}{N_{class}}$$

freq denotes the frequency of occurrence of a word, class: {pos, neg} Doing this for all our words in our vocabulary, we will get a table like this:

P(w_i | class)

```
+---------+------+------+
|  Word   | Pos  | Neg  |
+=========+======+======+
| i       | 0.24 | 0.25 |
| am      | 0.24 | 0.25 |
| happy   | 0.15 | 0.08 |
| because | 0.08 | 0.00 |
| learn   | 0.08 | 0.08 |
| nlp     | 0.08 | 0.08 |
| sad     | 0.08 | 0.17 |
| not     | 0.08 | 0.17 |
+---------+------+------+
| Sum     | 1    | 1    |
+---------+------+------+
```

In the Naive Bayes, We will find how each word is contributing to the sentiment, which can be calculated by the ratio of the probability of occurrence of the word for positive and negative class. Let's take an example; We can see that the probability of occurrence of the word 'sad' is more for negative than positive class. So, we will find the ratio of these probabilities for every word by the formula:

$$\frac{P(w_i|pos)}{P(w_i|neg)}$$

$$0 \qquad 1 \qquad \infty$$

Negative   Neutral   Positive
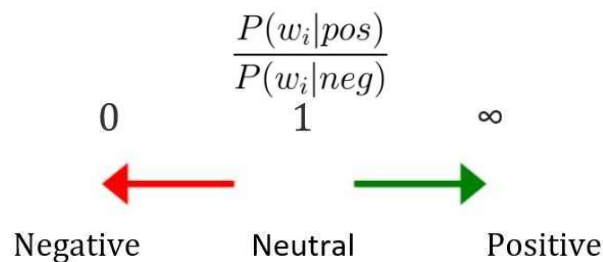
This ratio is known as the likelihood, and its value lies between $(0, \infty)$. The value tending to zero indicates that it has very low probability to occur in a positive tweet as compared to the probability to occur in a negative tweet and the ratio value tending to infinity shows that it has very low probability to occur in a negative tweet as compared to the probability to occur in a positive tweet. In other words, the high value of ratio implies positivity. Also, the ratio value 1 means that the name is neutral.

Laplace Smoothing

Some words might have occurred in any particular class only. The words which did not occur in the negative class will have probability 0 which makes the ratio undefined. So, we will use the Laplace smoothing technique to pursue this kind of situation. Let's take on how equation changes on applying Laplace smoothing:

$$P(w_i \mid class) = \frac{freq(w_i, class) + 1}{Nclass + V}$$

$Nclass$ = Frequency of all words in class

V = number of unique words in the vocabulary.

By adding '1' in the numerator makes the probability non zero. This factor is called alpha-factor and is between $(0,1]$; specifically, when we set this alpha-factor to 1, the smoothing is termed as Laplace smoothing. Also, the sum of probabilities will remain at 1.

Here in our example, the number of unique words is eight gives us V= 8.

After Laplace smoothing the table of the probability will look like this:

P(w_i | class)

```
+---------+------+------+
|  Word   | Pos  | Neg  |
+=========+======+======+
| i       | 0.19 | 0.20 |
| am      | 0.19 | 0.20 |
| happy   | 0.14 | 0.10 |
| because | 0.10 | 0.05 |
| learn   | 0.10 | 0.10 |
| nlp     | 0.10 | 0.10 |
| sad     | 0.10 | 0.15 |
| not     | 0.10 | 0.15 |
+---------+------+------+
| Sum     | 1    | 1    |
+---------+------+------+
```

Naive Bayes:

To estimate the sentiment of a tweet, we will take the product of the probability ratio of each word occurred in the tweet. Note, the words which are not present in our vocabulary will not contribute and will be taken as neutral. The equation for naive Bayes in our application will be like this:

$$\prod_{i=1}^{m} \frac{P(w_i|pos)}{P(w_i|neg)}$$

m = number of words in a tweet, w = set of words in a tweet

Since the data can be imbalanced and can cause biased results for a particular class, we multiply the above equation with a prior factor, which is the ratio of the probability of positive tweets to the probability of negative tweets.

$$\frac{P(pos)}{P(neg)} \prod_{i=1}^{m} \frac{P(w_i|pos)}{P(w_i|neg)}$$

complete equation of Naive Bayes Since we are taking the product of all these ratios, we can end up with a number too large or too small to be stored on our device, so here comes the concept of log-likelihood. We take the log over our equation of Naive Bayes.

$$log\left(\frac{P(pos)}{P(neg)}\prod_{i=1}^{n}\frac{P(w_i|pos)}{P(w_i|neg)}\right) \Rightarrow \quad log\frac{P(pos)}{P(neg)}+\sum_{i=1}^{n}log\frac{P(w_i|pos)}{P(w_i|neg)}$$

log prior + log likelihood

After taking the log of the likelihood equation, the scale will be changed as follows:

$-\infty$ 0 ∞

Negative   **Neutral**   **Positive**
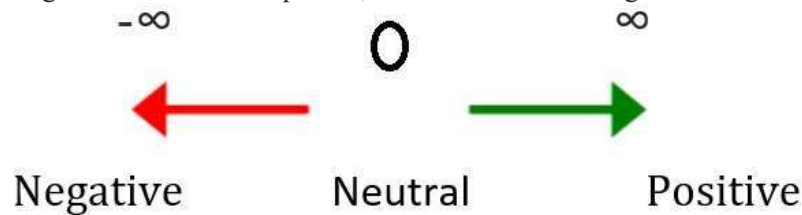
Stock price prediction

Stock prediction is the act of forecasting stock prices based on historical data. I used historical data in machine learning to recognize trends and understand the current market. Machine learning automates the trading process by using statistical models to draw insights and make predictions. Machine learning can collect and test a large amount of data, both structured and unstructured. It can apply suitable algorithms, transform, search for patterns, and make decisions based on the new data.

TESTING

The testing phase is an important part of software development. It is the Information zed system will help in automate process of finding errors and missing operations and also a complete verification to determine whether the objectives are met and the user requirements are satisfied. Software testing is carried out in three steps:
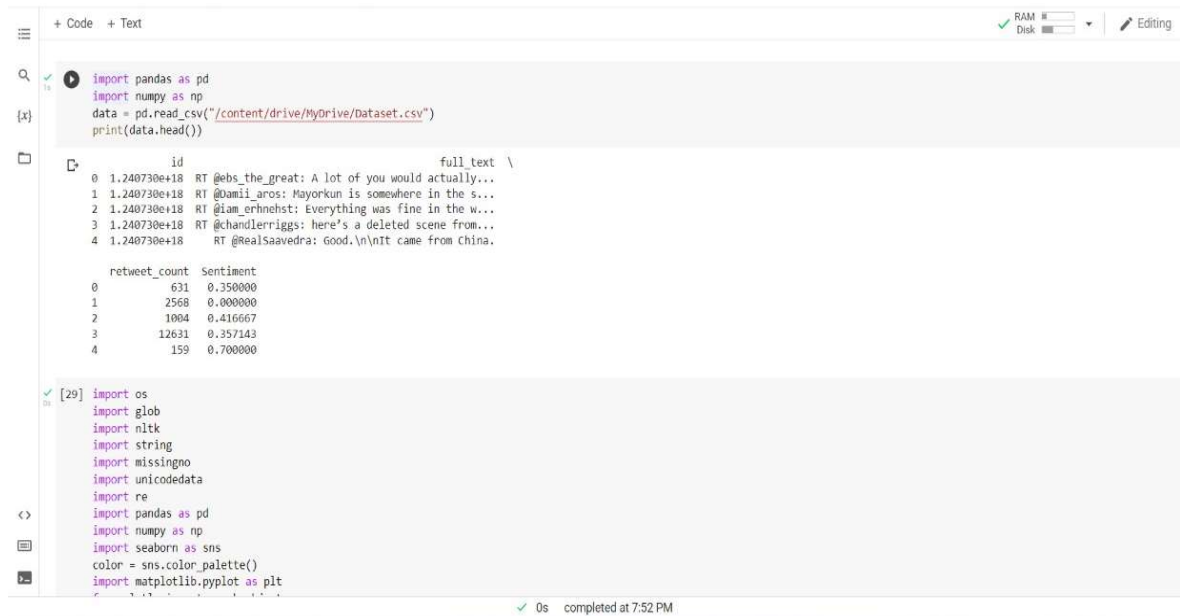
1. The first includes unit testing, where in each module is tested to provide its correctness, validity and also determine any missing operations and to verify whether the objective shave been met. Errors are noted down and corrected immediately.

2. Unit testing is the important and major part of the project. So errors are rectified easily in particular module and program clarity is increased. In this project entire system is divided into several modules and is developed individually. So unit testing is conducted to individual modules.

3. The second step includes Integration testing. It need not be the case, the software whose modules when run individually and showing perfect results, will also show perfect results when run as a whole.

```
        ┌──────────────┐
        │ fetch & store│
        │  the data    │
        └──────────────┘
               │
        ┌──────────────┐
        │ prepare the  │
        │ data for the │
        │  ML model    │
        └──────────────┘
               │
  ┌─────────┐  ◇  ┌─────────┐
  │Train the│←Divide→│Make     │
  │model on │ the data│predictions│
  │training │ into... │on test  │
  └─────────┘        └─────────┘
               │
        ┌──────────────┐
        │ predict and  │
        │ calculate    │
        │ metric perf. │
        └──────────────┘
               │
        ┌──────────────┐
        │Plot predictions│
        │ and display  │
        └──────────────┘
               │
        ┌──────────────┐
        │ read graph   │
        │ and perform  │
        │ buy,sell,hold│
        └──────────────┘
```
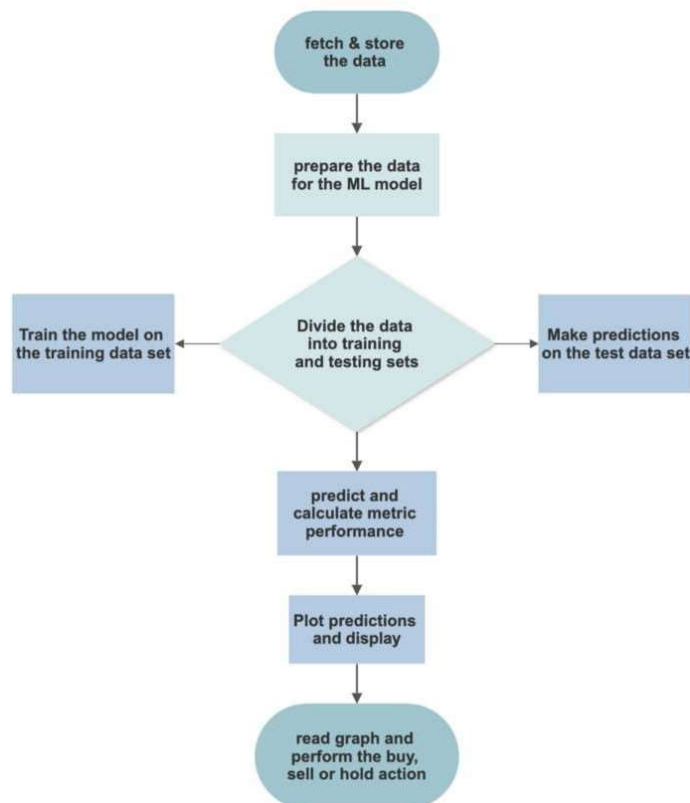
CHAPTER -8

SNAPSHOTS

SNAPSHOTS


```python
import pandas as pd
import numpy as np
data = pd.read_csv("/content/drive/MyDrive/Dataset.csv")
print(data.head())
```

```
          id                                         full_text  \
0  1.240730e+18  RT @ebs_the_great: A lot of you would actually...
1  1.240730e+18  RT @Damii_aros: Mayorkun is somewhere in the s...
2  1.240730e+18  RT @iam_erhnehst: Everything was fine in the w...
3  1.240730e+18  RT @chandlerriggs: here's a deleted scene from...
4  1.240730e+18      RT @RealSaavedra: Good.\n\nIt came from China.

   retweet_count  Sentiment
0            631   0.350000
1           2568   0.000000
2           1004   0.416667
3          12631   0.357143
4            159   0.700000
```

```python
import os
import glob
import nltk
import string
import missingno
import unicodedata
import re
import pandas as pd
import numpy as np
import seaborn as sns
color = sns.color_palette()
import matplotlib.pyplot as plt
```

✓ 0s   completed at 7:52 PM

```python
import os
import glob
import nltk
import string
import missingno
import unicodedata
import re
import pandas as pd
import numpy as np
import seaborn as sns
color = sns.color_palette()
import matplotlib.pyplot as plt
from plotly import graph_objects as go
nltk.download('stopwords')
from nltk.corpus import stopwords
eng_stopwords = set(stopwords.words("english"))
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
In [ ]:   # Plotting Bar diagram indicating Sentiment with Sentiment Score
          Path = "/content/drive/My Drive/IBM_Hackathon_2020/Final_Datasets/*.csv"
          csv_list = glob.glob(Path)# collecting all files having same path
          mylist=csv_list[1:4]
          for f in mylist:
            df=pd.read_csv(f)#reading the csv file
            data=df.Sentiment.value_counts()#Creating the dataframe of Sentiment values and its count
            base = os.path.basename(f)#name of the file in the path
            plt.figure(figsize=(10,10))
            plt.xlabel("Sentiment",fontsize=12)
            plt.ylabel("Sentiment Score",fontsize=12)
            plt.title("Sentiment Levels on "+str(base)+" dataset")
            sns.barplot(data.index,data.values,alpha=0.8)
```


Sentiment Levels on Train_Data.csv dataset
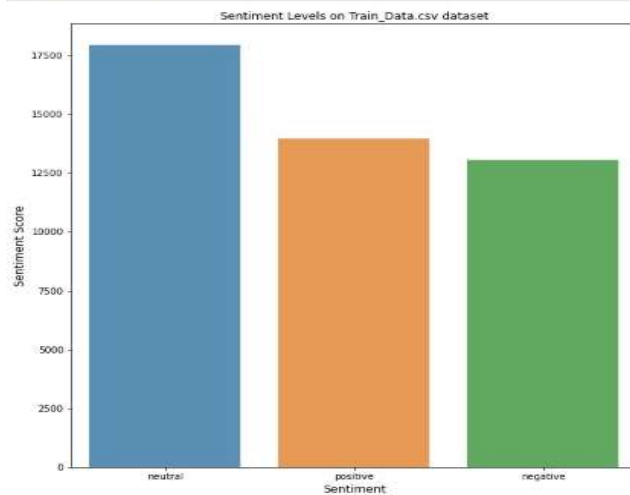
```
for f in mylist:
    df=pd.read_csv(f)
    data=df.Sentiment.value_counts()
    base = os.path.basename(f)
    fig = go.Figure(go.Funnelarea(
        values = data.values, text = ["Neutral","Positive","Negative"],
        marker = {"colors": ["deepskyblue", "lightsalmon", "tan"],

              "line": {"color": ["wheat", "wheat", "wheat"]}},
                title = {"position": "top center", "text": "Sentiment levels on "+str(base)+" dataset"}))
    fig.show()
```

```
In [ ]:  #Plotting the bar graph of top frequently 10 occuring words for each Sentiment in each datset
         for f in mylist:
             df=pd.read_csv(f)
             base = os.path.basename(f)
             data=df.Sentiment.value_counts()
             Analysis_Data = df
             Analysis_Data["full_text"]= Analysis_Data["full_text"].str.lower()#converting the text into lowercase
             Analysis_Data['full_text'] = Analysis_Data['full_text'].apply(lambda x: ' '.join([word for word in x.split() if word not in (eng_stopwords)]))#
             Analysis_Data['full_text'] = Analysis_Data['full_text'].str.replace("[^\w\s]","")#removing the emojis
             for i in range(0,3):
                 Sentiment = Analysis_Data[Analysis_Data['Sentiment'] == data.index[i]]#Creating the dataframe of having same sentiment
                 Word_frequency = pd.Series(' '.join(Sentiment.full_text).split()).value_counts()[:10]#Calculating the words frequency
                 plt.figure(figsize=(18,8))
                 sns.barplot(Word_frequency.index, Word_frequency.values, alpha=0.8)
                 plt.ylabel('Frequency', fontsize=12)
                 plt.xlabel('Top 10 Words appearing in Text of '+str(data.index[i])+' Sentiment Tweets in '+str(base)+' dataset', fontsize=12)
                 plt.show()
```


Top 10 Words appearing in Text of neutral Sentiment Tweets in Train_Data.csv dataset

```
[35]  Analysis_Data = pd.read_csv("/content/drive/MyDrive/Dataset.csv")
      Analysis_Data['full_text'] = Analysis_Data["full_text"].apply(lambda x: normalizeString(x)) #Normalising the string
      Analysis_Data["num_words"] = Analysis_Data["full_text"].apply(lambda x: len(str(x).split()))#calculating the number of words in each tweet
      Analysis_Data["num_unique_words"] = Analysis_Data['full_text'].apply(lambda x: len(set(str(x).split())))#Calculating the number of words
      Analysis_Data["num_chars"] = Analysis_Data['full_text'].apply(lambda x: len(str(x)))#Calculating the number of characters
      Analysis_Data["num_stopwords"] = Analysis_Data['full_text'].apply(lambda x: len([w for w in str(x).lower().split() if w in eng_stopwords]))#Calculating the number of stop words
      Analysis_Data["num_punctuations"] = Analysis_Data['full_text'].apply(lambda x: len([c for c in str(x) if c in string.punctuation]) )
      Analysis_Data["num_words_upper"] = Analysis_Data['full_text'].apply(lambda x: len([w for w in str(x).split() if w.isupper()]))
      Analysis_Data["mean_word_len"] = Analysis_Data['full_text'].apply(lambda x: np.mean([len(w) for w in str(x).split()]))
```

Analysis_Data.describe()

1 to 8 of 8 entries   Filter

index	id	retweet_count	Sentiment	num_words	num_unique_words	num_chars	num_stopwords	num_punctuations	num_words_upper	mean_w
count	602681.0	602681.0	602681.0	602681.0	602681.0	602681.0	602681.0	602681.0	602681.0	
mean	1.2407911915424576e+18	22384.469581752204	-0.0014419387254119506	20.151887980540288	18.418513276509465	107.53706853210903	7.39417038200972	1.3425145309044089	0.0	4.438352
std	3985207021481.49	33945.77830739983	0.3169507641163046	9.38805524911093	7.767575766293062	49.114941860616014	4.839168873162991	1.6331730423269213	0.0	0.7380948
min	1.24073e+18	0.0	-1.0	2.0	2.0	7.0	0.0	0.0	0.0	1.3265306
25%	1.24075e+18	2.0	-0.108333333	13.0	12.0	66.0	3.0	0.0	0.0	3.9285714
50%	1.24079e+18	1262.0	0.0	21.0	20.0	118.0	7.0	1.0	0.0	
75%	1.24083e+18	47400.0	0.136363636	25.0	23.0	136.0	10.0	2.0	0.0	4.909090
max	1.24086e+18	292989.0	1.0	122.0	107.0	897.0	46.0	91.0	0.0	15.555555

Show 10 v per page
Like what you see? Visit the data table notebook to learn more about interactive tables.

Double-click (or enter) to edit

✓ 0s   completed at 7:52 PM

CHAPTER- 9
CONCLUSION

CONCLUSION AND FUTURE WORK

we found that the most suitable algorithm for predicting the market price of a stock based on various data points from the historical data is the random forest algorithm. The algorithm will be a great asset for brokers and investors for investing money in the stock market since it is trained on a huge collection of historical data and has been chosen after being tested on a sample data. The project demonstrates the machine learning model to predict the stock value with more accuracy as compared to previously implemented machine learning models.

Future scope of this project will involve adding more parameters and factors like the financial ratios, multiple instances, etc. The more the parameters are taken into account more will be the accuracy. The algorithms can also be applied for analyzing the contents of public comments and thus determine patterns/relationships between the customer and the corporate employee. The use of traditional algorithms and data mining techniques can also help predict the corporation's performance .

CHAPTER - 10
REFERENCES

REFERENCES

- Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of Stock Market Prediction Using Machine Learning

- Approach", ICECA 2017.

- Loke.K.S. "Impact Of Financial Ratios And Technical Analysis On Stock Price Prediction Using Random Forests",

- IEEE,2017.

- Xi Zhang1, Siyu Qu1, Jieyun Huang1, Binxing Fang1, Philip Yu2, "Stock Market Prediction via Multi-Source Multiple

- Instance Learning." IEEE2018.

- VivekKanade, BhausahebDevikar, SayaliPhadatare, PranaliMunde, ShubhangiSonone. "Stock Market Prediction: Using

- Historical Data Analysis", IJARCSSE 2017.

- SachinSampatPatil, Prof. Kailash Patidar, Asst. Prof. Megha Jain, "A Survey on Stock Market Prediction Using SVM",

- IJCTET2016.

- https://www.cs.princeton.edu/sites/default/files/uploads/Saahil_magde.pdf

- Hakob GRIGORYAN, "A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent

- Component Analysis (ICA)", DSJ 2016.

- RautSushrut Deepak, ShindeIshaUday, Dr. D. Malathi, "Machine Learning Approach In StockMarket

- Prediction", IJPAM2017.

- Pei-Yuan Zhou , Keith C.C. Chan, Member, IEEE, and Carol XiaojuanOu, "Corporate Communication Network and

- Stock Price Movements: Insights From Data Mining", IEEE201