Prediction of stock prices using twitter sentiment analysis

A PROJECT REPORT

Submitted by

Ayush Singh [Reg No: RA1911003030415]
Priansh Gangrade [Reg No: RA1911003030394]
Shiva Maurya [Reg No: RA1911003030414]

Jatin Kesharwani [Reg No: RA1911003030412]

Under the guidance of **Mr. Karthick S.**
(Assistant Professor, Department of Computer Science and Engineering)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

Of

FACULTY OF ENGINEERING AND TECHNOLOGY

INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

SRM INSTITUTE OF SCIENCE AND TECNOLOGY

MAY 2023

BONAFIDE CERTIFICATE

This is to certify that Project Report “Prediction of Stock Prices using Twitter Sentiment Analysis ”, which is submitted by Ayush Singh(RA1911003030415), Priansh Gangrade(RA1911003030394), Jatin Kesharwani (RA1911003030412) and Shiva Maurya (RA1911003030414) in the partial fulfillment of the requirement for the award of degree B.Tech (CSE) of SRM Institute of Science and Technology, Delhi-NCR Campus, Modinagar, Ghaziabad is a record of the candidates’ own work carried out by them under my own supervision.

………………………… …………………………

Dr. Akash Punhani Mr. Karthick S.

HOD(CSE) Assistant Professor(CSE)

INTERNAL EXAMINER EXTERNAL EXAMINER

ABSTRACT

This project investigates the use of Twitter sentiment Analysis for predicting stock prices. The study utilizes a random forest Machine Learning Model to analyse sentiment features extracted from Twitter data and predict stock prices. The project explores the impact of different features and parameters on the accuracy of the model, including the selection of sentiment lexicons, the number of features used, and the time window of the stock price data.

Since the current crises that has inevitably impacted the financial market, market prediction has become more crucial than ever. The question of how risk managers can more accurately predict the evolution of their portfolio, while taking into consideration systemic risks brought on by a systemic crisis, is raised by the low rate of success of portfolio risk-management models. Sentiment analysis on natural language sentences can increase the accuracy of market prediction because financial markets are influenced by investor sentiments. Many investors also base their decisions on information taken from newspapers or on their instincts.

The project uses a dataset of over a thousand tweets related to the stocks of companies listed in time series from all sectors. Sentiment features are extracted using three different sentiment lexicons, and the feature selection process is performed using a chi-squared test. The random forest model is trained on a portion of the data and evaluated on a separate testing set. The results show that sentiment analysis can be used to predict stock prices.

In conclusion, the study demonstrates the potential of Twitter sentiment analysis for predicting stock prices, and the effectiveness of a random forest and support vector model for this task. The results highlight the importance of feature selection and parameter tuning in achieving accurate predictions and suggest that sentiment analysis could be a valuable addition to traditional stock market analysis methods.

ACKNOWLEDGEMENT

We would like to convey our heartfelt appreciation to our guide, Mr. Karthick S, for his invaluable advice, persistent encouragement, personal caring, prompt assistance, and for providing us with a wonderful environment in which to do research. Throughout the project, despite his hectic schedule, he has been cheery and polite in his support of our efforts to complete this research. We would also like to thank the faculty and students in our department for their assistance during our study. Finally, we'd like to thank our parents, family members, and friends for their unending love, support, and encouragement.

We feel compelled to articulate our thankfulness to Dr. Jitendra Singh and Dr. Rakesh Kumar Yadav, Project Coordinators, SRM Institute of Science and Technology, Delhi-NCR Campus, Modinagar, for there enlightening dexterity which was truly a perennial source of inspiration.

Ayush Singh

Priansh Gangrade

Shiva Maurya

Jatin Kesharwani

DECLARATION

We Ayush Singh(RA1911003030415), Priansh Gangrade(RA1911003030394), Jatin Kesharwani(RA1911003030412) and Shiva Maurya(RA1911003030414) hereby declare that the work which is being presented in the project report " Prediction of Stock Prices using Twitter Sentiment Analysis " is the record of authentic work carried out by us during the period from January 23 to May 23 and submitted by us in partial fulfillment for the award of the degree " Bachelor of Technology in Computer Science and Engineering" to SRM IST, NCR Campus, Ghaziabad (U. P.). This work has not been submitted to any other University or Institute for the award of and Degree/Diploma.

Ayush Singh (RA1911003030415)

Priansh Gangrade
(RA1911003030394)

Jatin Kesharwani
(RA1911003030412)

Shiva Maurya (RA1911003030414)

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: Introduction

The relationship between social media sentiment and stock prices has been a topic of interest among researchers and investors for several years. In recent times, Twitter has emerged as a popular platform for investors to express their opinions and sentiments about different companies, making it a valuable source of information for predicting stock prices. This research paper aims to explore the potential of using Twitter sentiment analysis to predict stock prices using two different machine learning models: Random Forest and Support Vector Machines (SVM).

In this paper, we first provide a review of the relevant literature on the relationship between social media sentiment and stock prices. We then explain the methodology used to collect and analyze Twitter data, including the pre-processing steps to clean and prepare the data for analysis. Next, we discuss the two machine learning models used in the study and their respective performances in predicting stock prices.

We present our findings based on the analysis of the data and the results obtained from the machine learning models. We also provide a comparison of the performance of the two models and their respective strengths and weaknesses in predicting stock prices using Twitter sentiment data. Finally, we conclude the paper by discussing the implications of our findings for investors and highlighting the potential of using Twitter sentiment analysis as a tool for predicting stock prices.

The primary objective of this research paper is to contribute to the existing body of literature on the relationship between social media sentiment and stock prices and to evaluate the effectiveness of two commonly used machine learning models in predicting stock prices using Twitter sentiment data. The paper also aims to provide insights and recommendations for investors looking to leverage social media sentiment analysis for better stock trading decisions.

CHAPTER 2: LITERATURE SURVEY

S.no	Topic	Members	Inference	Advantages	Disadvantages
1	Twitter mood predicts the stock market	Bollen, Mao, and Zeng (2011)	The study found that Twitter mood can predict the movement of the Dow Jones Industrial Average (DJIA) up to six days in advance. Specifically, the study found that an increase in the level of anxiety and concern expressed on Twitter predicted a subsequent decrease in the DJIA, while an increase in the level of positive mood expressed on Twitter predicted a subsequent increase in the DJIA.	The study provides evidence for the potential usefulness of social media data in predicting stock market movements. This could have important implications for investors and financial analysts who are interested in using social media data to inform their investment decisions.	The study has been criticized for its methodology, including the use of a small sample size and a relatively short time frame for analysis. Additionally, the study relied on automated sentiment analysis tools to analyze Twitter data, which may not accurately capture the nuances of human emotion and could lead to errors in predicting market movements. Finally, the study only focused on the DJIA and may not be applicable to other stock market indices
2	Comparing Twitter and Traditional Media Using Topic Models	Zhang, Skiena, and Sun (2011)	The study found that Twitter and traditional media cover different topics and have different levels of bias. Specifically, Twitter was found to be more focused on entertainment and social issues, while traditional media was more focused on politics and international news. The study also found that traditional media was more biased towards a liberal perspective, while Twitter was more evenly split between liberal and	The study provides insights into the differences between Twitter and traditional media in terms of content and bias. This could be useful for individuals and organizations interested in understanding the differences between these two types of media and how	The study focused on a specific time period and may not be generalizable to other time periods or contexts. Additionally, the study relied on topic modeling techniques, which may not accurately capture the full range of content and biases present in both Twitter and traditional media.

			conservative perspectives.	they might be used for different purposes.	Finally, the study did not explore the reasons behind the observed differences between Twitter and traditional media, leaving open questions about the underlying causes of these differences.
3	The impact of social media on consumer purchase intention: The moderating role of social capital	Yeh and Li (2013)	The study found that social media has a significant positive impact on consumer purchase intention. Furthermore, social capital, which refers to the resources that individuals have access to through their social networks, moderates the relationship between social media and consumer purchase intention. Specifically, the study found that the positive effect of social media on consumer purchase intention is stronger for individuals with higher levels of social capital.	The study provides insights into the role of social media in influencing consumer behavior, which has important implications for marketers and businesses. The study also highlights the importance of social capital in moderating the impact of social media on consumer purchase intention, which could help businesses target their marketing efforts more effectively	The study was conducted using a cross-sectional survey design, which limits the ability to establish causality. Additionally, the study was limited to a single country (China) and may not be generalizable to other cultural contexts. Finally, the study relied on self-reported measures, which may be subject to response bias.
4	Tweeting to Feel Connected: A Model of Social Media Use in Narcissism	Sprenger, Welpe, and Gloor (2014)	The study found that individuals high in narcissism are more likely to use Twitter to seek attention and feel connected to others. However, this behavior can also lead to negative outcomes, such as conflict and aggression on Twitter.	The study provides insights into the role of social media in the behavior of individuals with high levels of narcissism. This could be useful for mental health professionals, as well as individuals who are interested in understanding their own social	The study relied on self-reported measures of narcissism and social media use, which may be subject to response bias. Additionally, the study only focused on Twitter and may not be applicable to other social media platforms. Finally, the study did not explore the potential

				media use and how it may be influenced by personality traits.	positive outcomes of social media use for individuals with high levels of narcissism, leaving open questions about the potential benefits of this behavior.
5	Big data: A survey	Mao, Wei, and Liu (2015)	The study highlights the importance of big data in today's society, including its potential to revolutionize various industries, such as healthcare, finance, and marketing. The study also identifies several challenges associated with big data, including privacy concerns, data quality issues, and the need for new analytical methods to process and make sense of large data sets.	The study provides a comprehensive overview of big data, including its definition, applications, challenges, and future directions. This could be useful for individuals and organizations interested in understanding the potential benefits and limitations of big data, as well as the best practices for collecting, storing, and analyzing large data sets	The study does not provide original research findings, but rather synthesizes existing literature on big data. Additionally, the study may not be comprehensive enough to cover all aspects of big data, as the field is constantly evolving and new developments are emerging. Finally, the study may be limited by its publication date, as the field of big data is rapidly advancing and new research is constantly being published.
6	A review of affective computing: From unimodal analysis to multimodal fusion	Poria, Cambria, and Hussain (2015)	The study highlights the importance of affective computing in a wide range of applications, including healthcare, marketing, and education. The study also identifies the challenges associated with affective computing, including the need for more advanced algorithms and the importance of considering cultural and social context when analyzing emotions and sentiments.	The study provides a comprehensive overview of affective computing, including its history, current state of the art, and future directions. This could be useful for researchers and practitioners interested in using affective computing in their work, as well as for individuals who are interested in understanding the potential	The study does not provide original research findings, but rather synthesizes existing literature on affective computing. Additionally, the study may not be comprehensive enough to cover all aspects of affective computing, as the field is constantly evolving and new developments are emerging. Finally, the study may be limited by its publication date, as the field of

				benefits and limitations of this technology.	affective computing is rapidly advancing and new research is constantly being published.
7	Exploring the use of social media for event detection and earthquake reporting	Nguyen, Shirai, and Velcin (2015)	The study found that Twitter can be an effective tool for detecting and reporting earthquakes, often providing faster and more comprehensive coverage than traditional news sources. The study also found that the accuracy of Twitter-based earthquake reports can be improved by combining multiple sources of information, such as geolocation data and user credibility ratings.	The study provides insights into the potential of social media for detecting and reporting natural disasters, which could be useful for emergency responders and other organizations involved in disaster management. The study also highlights the importance of considering the credibility and reliability of social media sources when using them for event detection and reporting.	The study focused on a specific type of event (earthquakes) and may not be generalizable to other types of natural disasters or events. Additionally, the study relied on Twitter data, which may not be representative of all social media platforms or all users. Finally, the study did not explore the potential limitations or drawbacks of using social media for event detection and reporting, leaving open questions about the reliability and accuracy of these methods in different contexts.
8	Entity discovery and linking in social media	Ding, Zhang, and Liu (2016)	The study proposes a novel approach for entity discovery and linking in social media, which combines textual, structural, and semantic features to improve the accuracy and efficiency of the process. The study found that the proposed approach outperformed several existing methods in terms of both precision and recall.	The study provides a new approach to entity discovery and linking in social media, which could be useful for researchers and practitioners working with large-scale social media data. The proposed method takes into account multiple sources of information	The study was evaluated on a specific dataset and may not be generalizable to other types of data or social media platforms. Additionally, the proposed method may require significant computational resources and expertise to implement, making it less accessible to some users.

				to improve the accuracy and efficiency of the process, and could be adapted to work with different types of social media platforms and data sources.	Finally, the study did not compare the proposed method with other state-of-the-art approaches for entity discovery and linking, leaving open questions about the relative performance of different methods.
9	A survey of text mining in social media	Gomaa and Fahmy (2017)	The study found that text mining techniques are widely used in social media research, with applications in a range of domains including marketing, politics, and healthcare. The study also identified several challenges associated with text mining in social media, including the need for specialized techniques to handle the unique characteristics of social media data (such as the use of informal language and the presence of noise and ambiguity).	The study provides a useful overview of text mining techniques applied to social media data, including their strengths and limitations. The survey could be helpful for researchers and practitioners who are new to the field of text mining or who are interested in exploring new applications for social media data. The study also highlights the potential benefits of using text mining techniques to analyze social media data, such as the ability to identify emerging trends and to gain insights into public opinion and sentiment.	The study is based on a survey of existing literature and does not provide original research findings. Additionally, the survey may not be comprehensive enough to cover all relevant research on text mining in social media, as the field is constantly evolving and new research is emerging. Finally, the study does not provide a detailed analysis of the limitations and challenges associated with specific text mining techniques, leaving open questions about the reliability and validity of different methods in different contexts.
10	Sentiment analysis of social media for health care applications	Shah and Zadeh (2017)	The study found that sentiment analysis can be a useful tool for extracting and analyzing health-related information from social media data, including identifying patient experiences, opinions, and concerns. The study also identified several challenges	The study provides insights into the potential use of sentiment analysis techniques for health care applications, which could be useful for	The study was based on a limited dataset and may not be generalizable to other health-related contexts or social media platforms. Additionally, the study did not

			associated with using sentiment analysis in health care applications, including the need for specialized algorithms and tools to handle the unique characteristics of health-related data (such as medical terminology and the presence of domain-specific jargon).	researchers and practitioners in the medical field. The study also highlights the potential benefits of using social media data to gain insights into patient experiences and opinions, which could inform health care policy and practice. The study is particularly relevant in the context of the growing trend of patients using social media to discuss their health-related experiences.	provide a detailed analysis of the limitations and challenges associated with specific sentiment analysis techniques, leaving open questions about the reliability and validity of different methods in different contexts. Finally, the study did not address potential ethical concerns associated with using social media data for health care research, such as issues of privacy and informed consent.
11	An overview of topic modeling and its current applications in bioinformatics	Yu, Wang, and Huang (2017)	: The study found that topic modeling techniques can be useful for identifying meaningful patterns and relationships in complex biological data, such as gene expression data and protein-protein interaction networks. The study also identified several challenges associated with using topic modeling in bioinformatics, including the need for specialized algorithms and tools to handle the high-dimensional and noisy nature of biological data.	The study provides a useful overview of topic modeling techniques and their potential applications in the field of bioinformatics, which could be helpful for researchers and practitioners who are new to the field or who are interested in exploring new applications for their data. The study also highlights the potential benefits of using topic modeling techniques to analyze biological data, such as the ability to identify novel biomarkers and drug targets.	The study may not be comprehensive enough to cover all relevant research on topic modeling in bioinformatics, as the field is constantly evolving and new research is emerging. Additionally, the study does not provide a detailed analysis of the limitations and challenges associated with specific topic modeling techniques, leaving open questions about the reliability and validity of different methods in different contexts. Finally, the study may be difficult to understand for

					readers who are not familiar with the field of bioinformatics or with the technical details of topic modeling.
12	Predicting crowdfunding success with linguistic and non-linguistic signals: A meta-analysis	Zhang, Fuehres, and Gloor (2018)	The study found that both linguistic and non-linguistic signals can be useful for predicting crowdfunding success. Specifically, linguistic signals such as sentiment, complexity, and emotional content were found to be strong predictors of campaign success, as were non-linguistic signals such as video length and number of images. The study also identified several limitations and challenges associated with using these signals, including the need for more accurate and reliable data, and the difficulty of generalizing results across different crowdfunding platforms and campaigns.	The study provides insights into the potential use of linguistic and non-linguistic signals in predicting crowdfunding success, which could be useful for researchers, practitioners, and entrepreneurs in the crowdfunding industry. The study also highlights the potential benefits of using a combination of linguistic and non-linguistic signals in predicting success, which could improve the accuracy and reliability of predictions.	The study was based on a meta-analysis of existing research, which may not be comprehensive enough to cover all relevant studies on the topic. Additionally, the study did not address potential ethical concerns associated with using linguistic and non-linguistic signals to predict success, such as issues of privacy and fairness. Finally, the study did not provide a detailed analysis of the limitations and challenges associated with specific linguistic and non-linguistic signals, leaving open questions about the reliability and validity of different methods in different contexts.
13	A topic modeling based analysis of social media data for public opinion sensing	Lv, Li, and Wang (2019)	The study found that topic modeling techniques can be used to extract and analyze public opinion on social media platforms, which can provide valuable insights for policymakers, researchers, and other stakeholders. Specifically, the study demonstrated the effectiveness of using Latent Dirichlet Allocation (LDA) to identify and analyze the key topics and sentiment expressed in	The study provides a useful example of how topic modeling techniques can be used to extract and analyze public opinion on social media, which could be valuable for researchers and practitioners working in	The study may not be generalizable to all social media platforms or populations, and the effectiveness of topic modeling techniques may depend on the specific issue being studied and the characteristics of the social media data being analyzed.

			social media data related to a specific issue. The study also identified several limitations and challenges associated with using topic modeling in this context, including the need for accurate and representative data, and the difficulty of generalizing results across different social media platforms and populations.	fields such as public opinion research, political science, and communication studies. The study also demonstrates the potential of using topic modeling to identify and analyze sentiment expressed in social media data, which could be useful for understanding public opinion on a range of issues.	Additionally, the study did not address potential ethical concerns associated with analyzing social media data, such as issues of privacy and consent. Finally, the study may be difficult to understand for readers who are not familiar with the technical details of topic modeling.

14	A review of sentiment analysis research in Chinese language	Wang and Chen (2020)	The study found that sentiment analysis research in the Chinese language has grown rapidly in recent years, and has focused on a range of topics including lexicon-based methods, machine learning techniques, and cross-lingual sentiment analysis. The study also identified several challenges and limitations associated with sentiment analysis in the Chinese language, including the lack of reliable and comprehensive lexicons, the difficulty of handling Chinese idioms and other linguistic features, and the need for more research on the cultural and linguistic differences that affect sentiment analysis across different languages.	The study provides a useful overview of sentiment analysis research in the Chinese language, which could be valuable for researchers, practitioners, and policymakers interested in analyzing sentiment in Chinese language data. The study also identifies several areas where further research is needed, which could guide future work in the field. Finally, the study highlights the importance of considering cultural and linguistic factors in sentiment analysis, which could improve the accuracy and reliability of sentiment analysis across different languages and cultures.	The study may be difficult for readers who are not familiar with sentiment analysis or the Chinese language, and may not be relevant to those who work primarily in other languages or domains. Additionally, the study is a literature review and does not present original research, so it may not be as informative as other studies that present new findings or data. Finally, the study may not address all of the potential challenges and limitations associated with sentiment analysis in the Chinese language, as the field is constantly evolving and new issues may arise.

CHAPTER 3: EXISTING PROBLEM AND PROPOSED SOLUTION

A key challenge in the stock market prediction using Twitter sentiment analysis is to analyse a large volume of real-time data and extract valuable information that can provide insights into the market trends. Therefore, the objective of this project is to explore the potential of using machine learning algorithms to analyse tweets related to specific companies or industries and predict their impact on the stock prices. The project aims to identify the most relevant Twitter features that can affect stock prices and develop a robust prediction model that can outperform traditional methods of stock market analysis. The proposed solution for the stock market prediction using Twitter sentiment analysis project is to develop a machine learning model that uses natural language processing techniques to analyse tweets related to specific companies or industries and predict their impact on stock prices. The solution involves the following steps:

1. Data Collection: Collecting relevant tweets related to the specific companies or industries from Twitter API.
2. Stock Information: Collect the relevant stock data, including stock prices, trading volumes, and any other relevant data.
3. Data Pre-processing: Preprocess the collected data by cleaning, filtering, and transforming it into a format that can be used by the model. For example, for the Twitter data, you may need to filter out irrelevant tweets, clean up the text data by removing stop words, hashtags, and mentions, and then convert the text data into numerical features using techniques such as word embedding
4. Feature Extraction: Extracting relevant features from the tweets, such as sentiment scores, keywords, and hashtags, that can influence the stock prices.
5. Data Integration: Integrate the stock data and Twitter data by aligning them based on date. You can use the date from the tweets and stock datato match them up.
6. Model Development: Developing a machine learning model that utilizes the extracted features to predict the stock prices.
7. Model Training: Train a random forest model on the integrated data to predict the stock price based on the Twitter sentiment features. Split the data into training and testing sets to evaluate the performance of the model.
8. Model Evaluation: Evaluating the model's performance using appropriate metrics such as accuracy, precision, recall, and F1 score and other relevant metrics. You can also plot the predicted stock prices against the actual stock prices to visualize the performance of the model..
9. Model Optimization: Optimizing the model to improve its accuracy and reliability.
10. Model Deployment: Deploy the model and use it to predict the stock price based on new Twitter data in real-time

Overall, this solution aims to provide investors with accurate and timely predictions of stock prices based on the sentiment analysis of Twitter data, enabling them to make informed investment decisions.

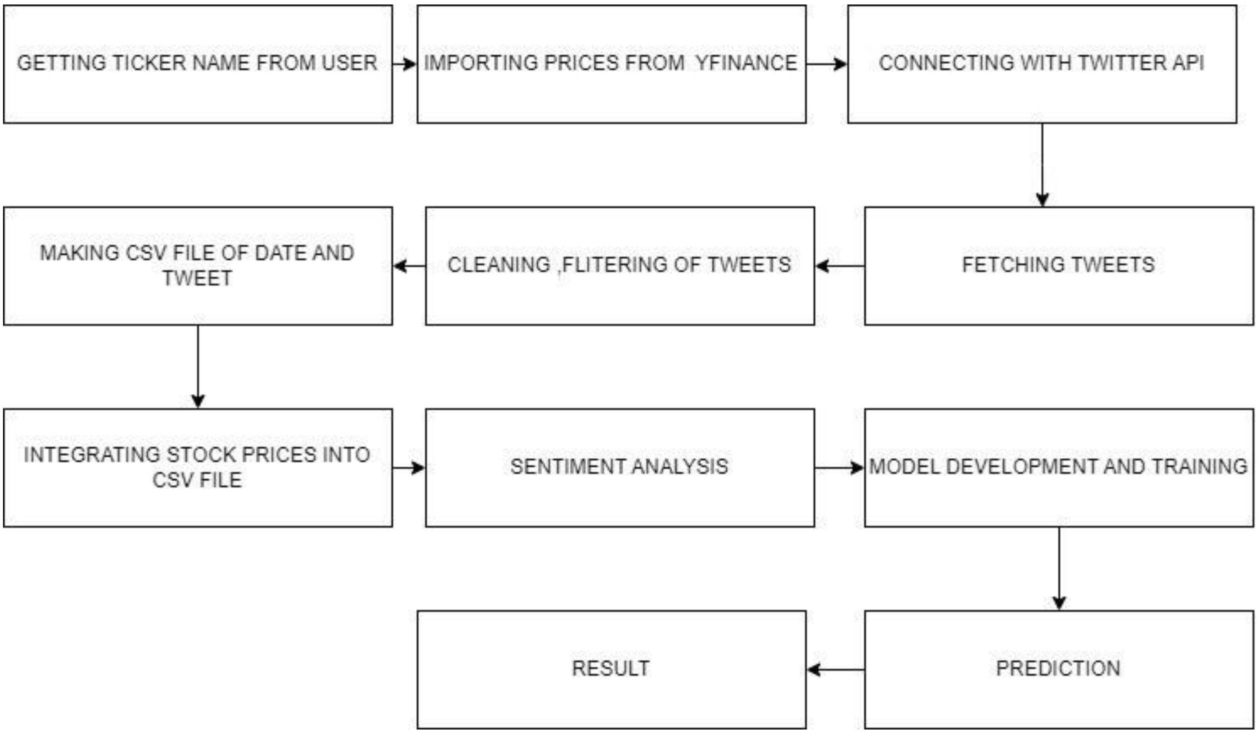
CHAPTER 4:METHODOLGIES

Tweet Extraction: The first process is to extract the tweets from twitter using TWEEPY. After the tweets are fetched from twitter, special characters are removed from those tweets. The tweets are then displayed with their corresponding dates in the form of a data frame.

Dataset: After the extraction of tweets, historical data of that particular company or commodity is downloaded from the Yahoo Finance website. Yahoo Finance is a website that provides live stock prices of the company or commodity and also provides downloadable csv files of the historical data.

Processing of Data: Price's column is then added to the data frame after the historical data is downloaded. Close Price of the company or the commodity is added to the Price column of the data frame. Some dates would not include any price due to some reasons like holidays. To fill in the values of the empty rows of the "Prices" column, mean of the available prices is determined and the empty rows are filled with this mean value.

Sentiment Analysis: Four new columns are added in the data frame. Comp, Negative, Neutral and Positive. Comp tells whether the sentence or the tweet is overall negative or positive. If the value of Comp is negative then, the sentence is negative and if the value of Comp is positive then, the sentence is positive


GETTING TICKER NAME FROM USER
IMPORTING PRICES FROM YFINANCE
CONNECTING WITH TWITTER API
FETCHING TWEETS
CLEANING ,FLITERING OF TWEETS
MAKING CSV FILE OF DATE AND TWEET
INTEGRATING STOCK PRICES INTO CSV FILE
SENTIMENT ANALYSIS
MODEL DEVELOPMENT AND TRAINING
PREDICTION
RESULT

Flowchart of the proposed system

4.1 RANDOM FOREST ALGORITHM

Random forest algorithm is a popular machine learning algorithm used for classification, regression, and other tasks. It is an ensemble method that combines multiple decision trees to make more accurate predictions.

The algorithm works by creating a set of decision trees, each trained on a different subset of the training data and using a random selection of features for each tree. During prediction, the new data point is passed through each decision tree, and the final prediction is made based on the average (for regression) or majority (for classification) vote of the individual tree predictions.

Random forest algorithm is known for its high accuracy, robustness, and ability to handle large datasets with high dimensionality. It is widely used in a variety of applications such as image classification, bioinformatics, and finance.

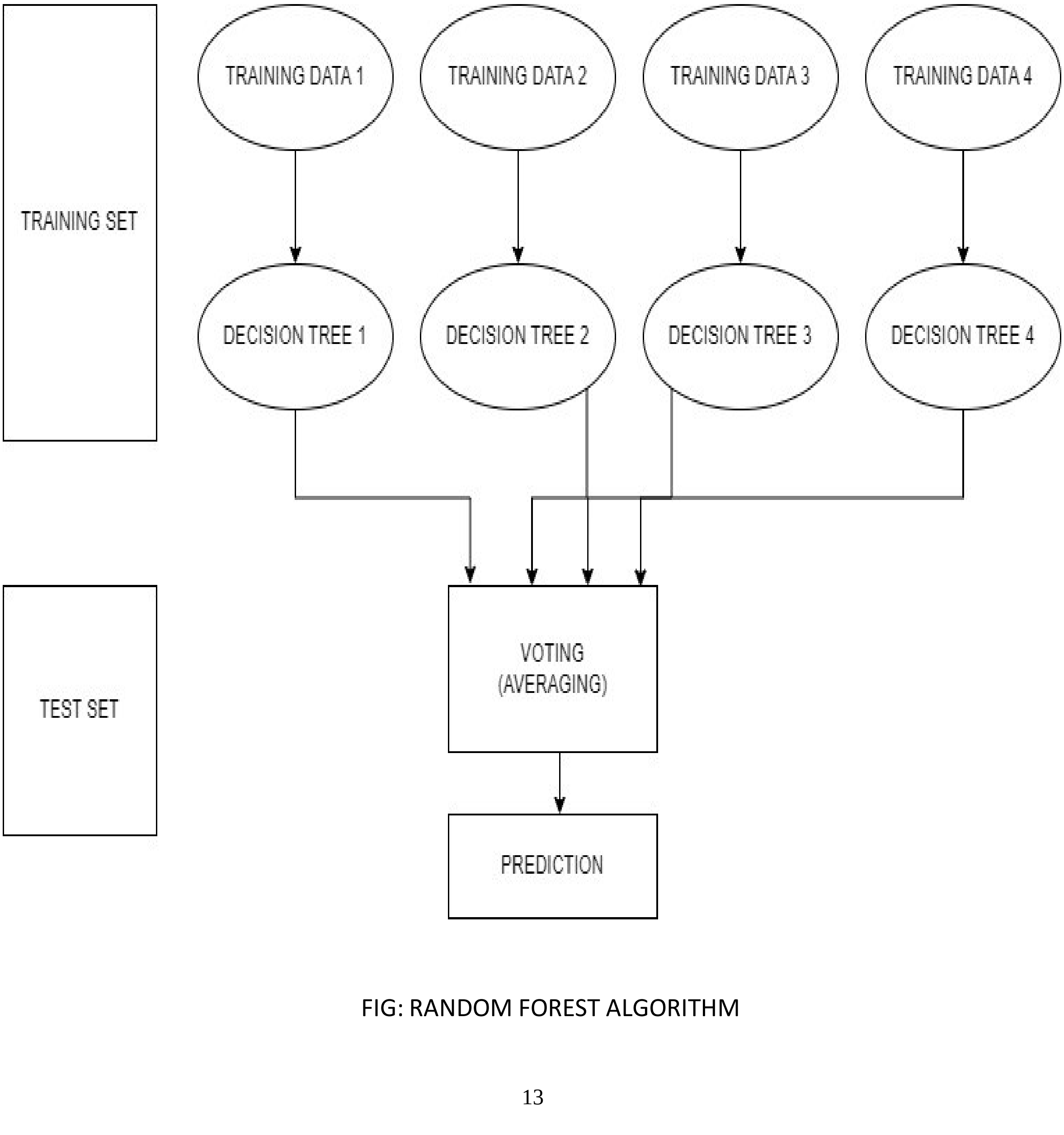


FIG: RANDOM FOREST ALGORITHM

4.2 SUPPORT VECTOR ALGORITHM

Support Vector Machine (SVM) is a popular supervised machine learning algorithm used for classification and regression tasks. It is based on the idea of finding a hyperplane in a high-dimensional space that separates the input data into two or more classes.

The SVM algorithm works by first mapping the input data into a higher-dimensional space using a kernel function. This is done to make it easier to find a hyperplane that can separate the data into different classes. The hyperplane is chosen in such a way that it maximizes the margin between the closest points of different classes. The points closest to the hyperplane are called support vectors.

The SVM algorithm tries to find the optimal hyperplane by solving an optimization problem. The objective is to maximize the margin between the support vectors while minimizing the classification error. The optimization problem can be solved using techniques such as quadratic programming or gradient descent.

Once the optimal hyperplane is found, new data points can be classified by mapping them into the same high-dimensional space and checking which side of the hyperplane they fall on. If a data point falls on the positive side of the hyperplane, it is classified as one class, and if it falls on the negative side, it is classified as the other class.

SVM is a powerful algorithm that can handle high-dimensional data, is effective in handling both linearly separable and non-linearly separable data, and has been widely used in applications such as imageclassification, text classification,and predictions.

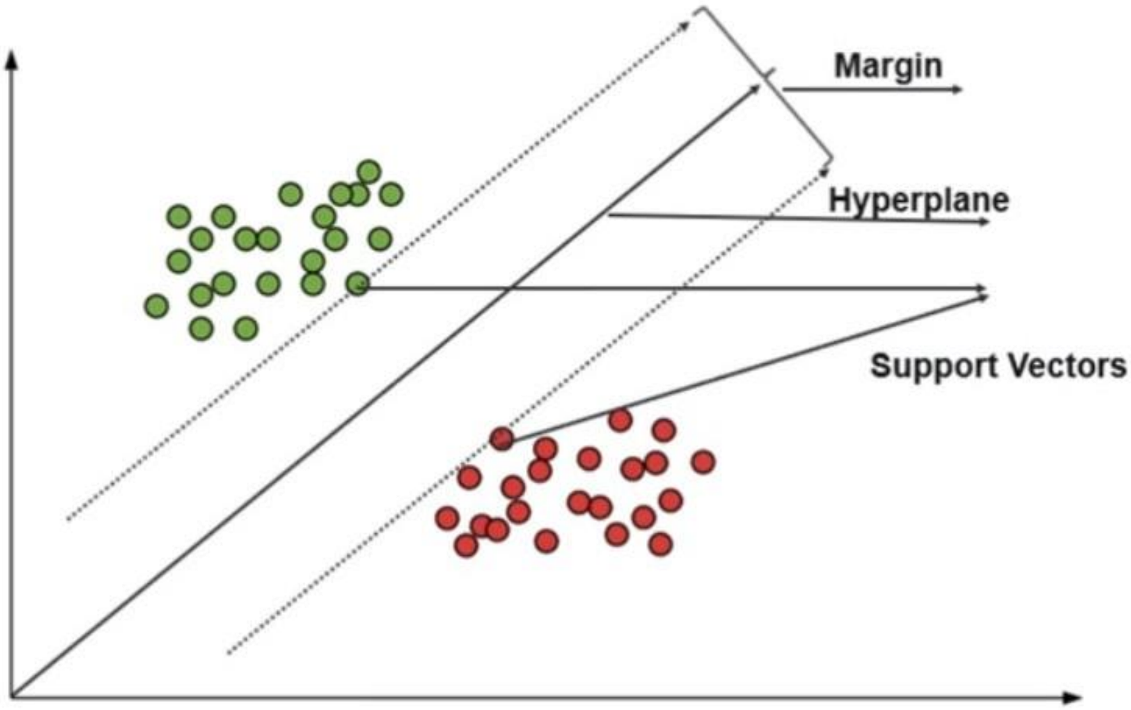


FIG: SUPPORT VECTOR ALGORITHM

CHAPTER 5 TOOLS USED

5.1 TWEEPY

Tweepy is a Python library for accessing and interacting with Twitter's API (Application Programming Interface). It provides a simple and convenient way to communicate with Twitter's API and perform various operations, such as searching for tweets, streaming real-time tweets, updating your status, and more.

Tweepy is widely used by developers and data scientists who want to work with Twitter data for research, marketing, and other purposes. It is built on top of Twitter's official API, which means that you can access all of the features and functions provided by Twitter's API using Tweepy.

Overall, Tweepy is a powerful and flexible library for accessing Twitter's API, and it has become a popular tool for working withTwitter data in Python.

5.2 PANDAS

Pandas is a Python library for data manipulation and analysis. It provides easy-to-use data structures and data analysis tools for working with structured data, such as tabular, time-series, and matrix data.

The core data structure in Pandas is the DataFrame, which is a two- dimensional table-like structure with rows and columns. The DataFrame can be thought of as a spreadsheet, where each row nrepresents a data record and each column represents a data attribute or feature.Pandas provides a wide range of tools for manipulating and analysing data in a Data Frame.

Pandas is widely used in data science, machine learning, and other fields that involve working with structured data. It provides a powerful and flexible set of tools for data manipulation and analysis, making itan essential library for any data scientist or analyst working with Python.

5.3 YFINANCE

yfinance (short for "Yahoo! Finance") is a Python library that allows you to download financial data from Yahoo! Finance. It provides an easy-to-use interface for accessing historical market data, as well as real-time data for stocks, ETFs, mutual funds, currencies, and more.

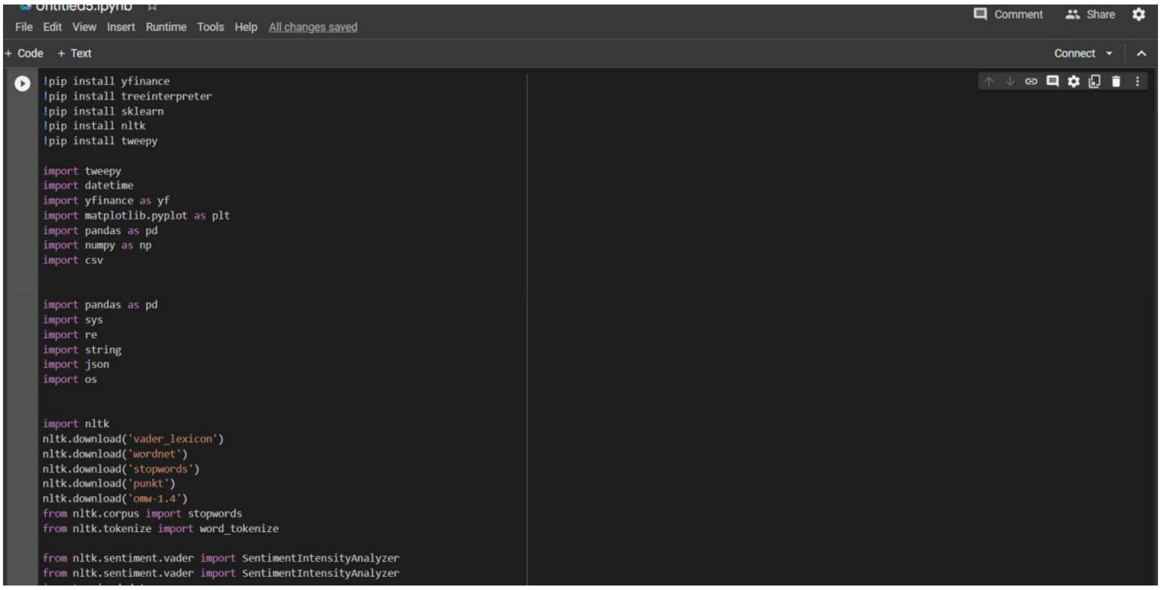
With yfinance, you can download historical price data for stocks and other financial instruments, as well as information about dividends, stock splits, and other corporate actions. You can also retrieve real-time stock prices and other market data, such as volume and market capitalization.

5.4 MATPLOTLIB

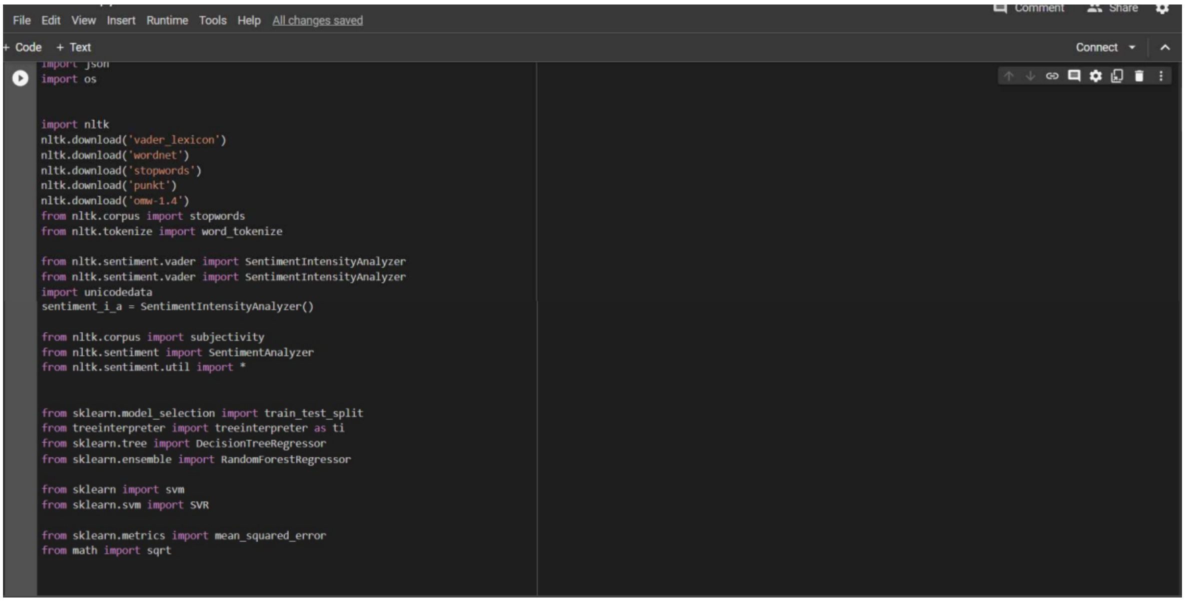
Matplotlib is a Python library for creating data visualizations, such as line plots, bar charts, scatter plots, histograms, and more. It provides a comprehensive set of tools for creating high-quality and publication- ready visualizations for scientific and engineering applications.

CHAPTER:6 CODING AND IMPLEMENTATION

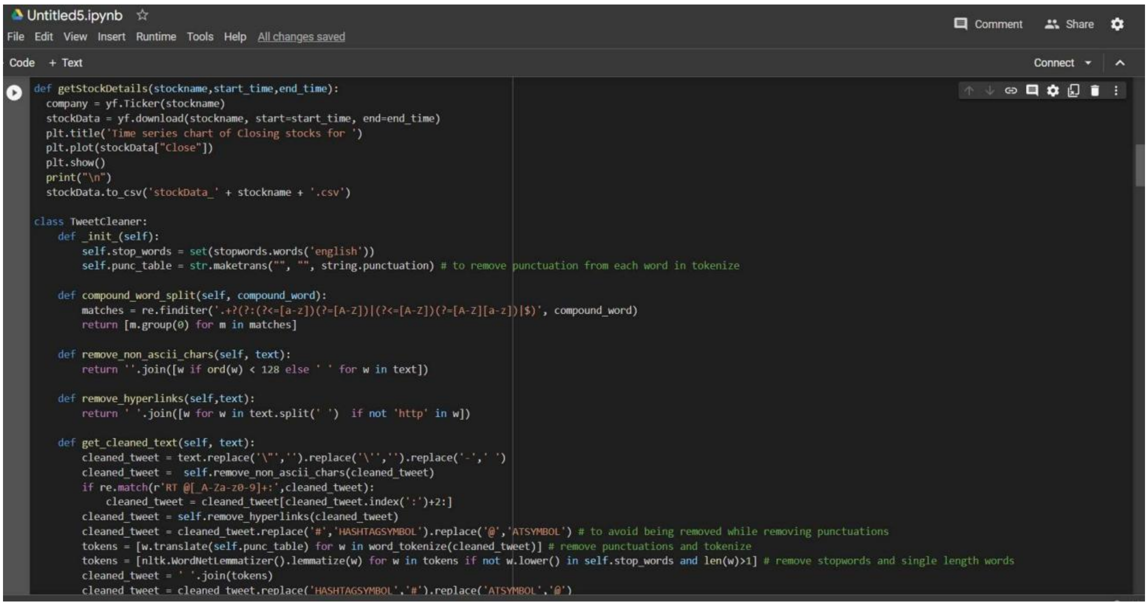
6.1 CODE



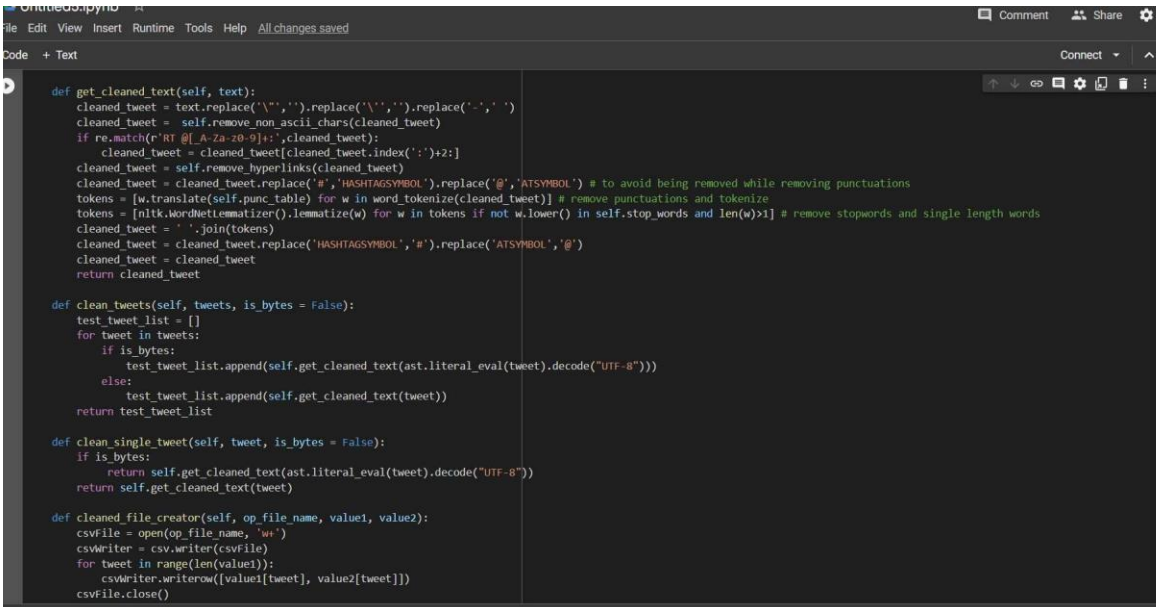
```
!pip install yfinance
!pip install treeinterpreter
!pip install sklearn
!pip install nltk
!pip install tweepy

import tweepy
import datetime
import yfinance as yf
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import csv


import pandas as pd
import sys
import re
import string
import json
import os


import nltk
nltk.download('vader_lexicon')
nltk.download('wordnet')
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('omw-1.4')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

from nltk.sentiment.vader import SentimentIntensityAnalyzer
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```



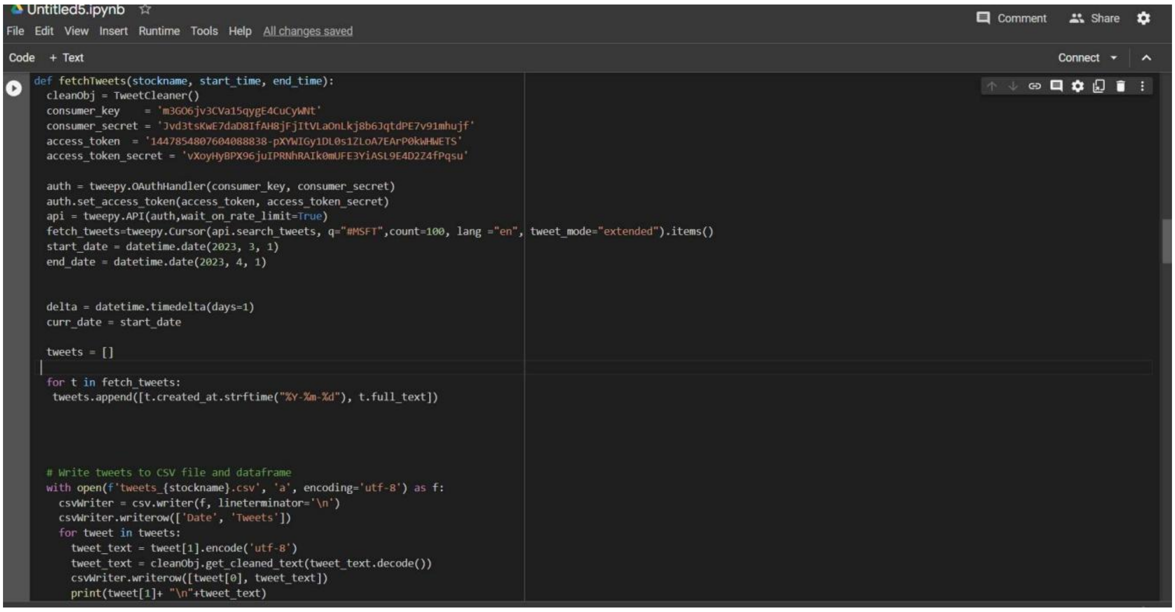
```
import json
import os


import nltk
nltk.download('vader_lexicon')
nltk.download('wordnet')
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('omw-1.4')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

from nltk.sentiment.vader import SentimentIntensityAnalyzer
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import unicodedata
sentiment_i_a = SentimentIntensityAnalyzer()

from nltk.corpus import subjectivity
from nltk.sentiment import SentimentAnalyzer
from nltk.sentiment.util import *


from sklearn.model_selection import train_test_split
from treeinterpreter import treeinterpreter as ti
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor

from sklearn import svm
from sklearn.svm import SVR

from sklearn.metrics import mean_squared_error
from math import sqrt
```



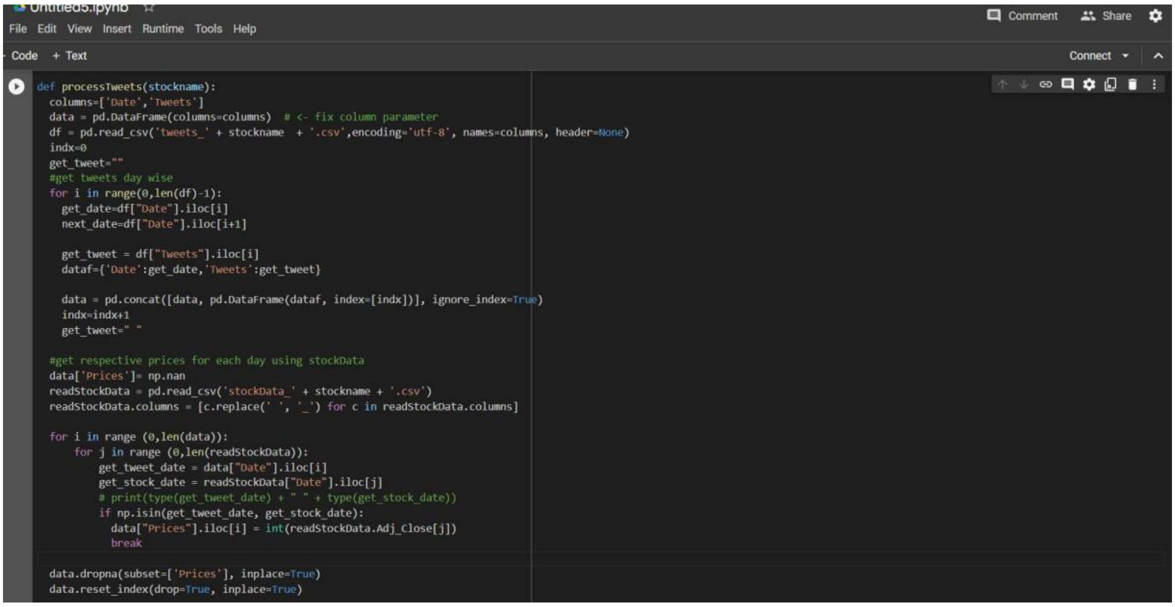
```
def getStockDetails(stockname,start_time,end_time):
  company = yf.Ticker(stockname)
  stockData = yf.download(stockname, start=start_time, end=end_time)
  plt.title('Time series chart of Closing stocks for ')
  plt.plot(stockData["Close"])
  plt.show()
  print("\n")
  stockData.to_csv('stockData_' + stockname + '.csv')

class TweetCleaner:
    def _init_(self):
        self.stop_words = set(stopwords.words('english'))
        self.punc_table = str.maketrans("", "", string.punctuation) # to remove punctuation from each word in tokenize

    def compound_word_split(self, compound_word):
        matches = re.finditer('.+?(?:(?<=[a-z])(?=[A-Z])|(?<=[A-Z])(?=[A-Z][a-z])|$)', compound_word)
        return [m.group(0) for m in matches]

    def remove_non_ascii_chars(self, text):
        return ''.join([w if ord(w) < 128 else ' ' for w in text])

    def remove_hyperlinks(self,text):
        return ' '.join([w for w in text.split(' ')  if not 'http' in w])

    def get_cleaned_text(self, text):
        cleaned_tweet = text.replace('\"','').replace('\'','').replace('-',' ')
        cleaned_tweet =  self.remove_non_ascii_chars(cleaned_tweet)
        if re.match(r'RT @[_A-Za-z0-9]+:',cleaned_tweet):
            cleaned_tweet = cleaned_tweet[cleaned_tweet.index(':')+2:]
        cleaned_tweet = self.remove_hyperlinks(cleaned_tweet)
        cleaned_tweet = cleaned_tweet.replace('#','HASHTAGSYMBOL').replace('@','ATSYMBOL') # to avoid being removed while removing punctuations
        tokens = [w.translate(self.punc_table) for w in word_tokenize(cleaned_tweet)] # remove punctuations and tokenize
        tokens = [nltk.WordNetLemmatizer().lemmatize(w) for w in tokens if not w.lower() in self.stop_words and len(w)>1] # remove stopwords and single length words
        cleaned_tweet = ' '.join(tokens)
        cleaned_tweet = cleaned_tweet.replace('HASHTAGSYMBOL','#').replace('ATSYMBOL','@')
```



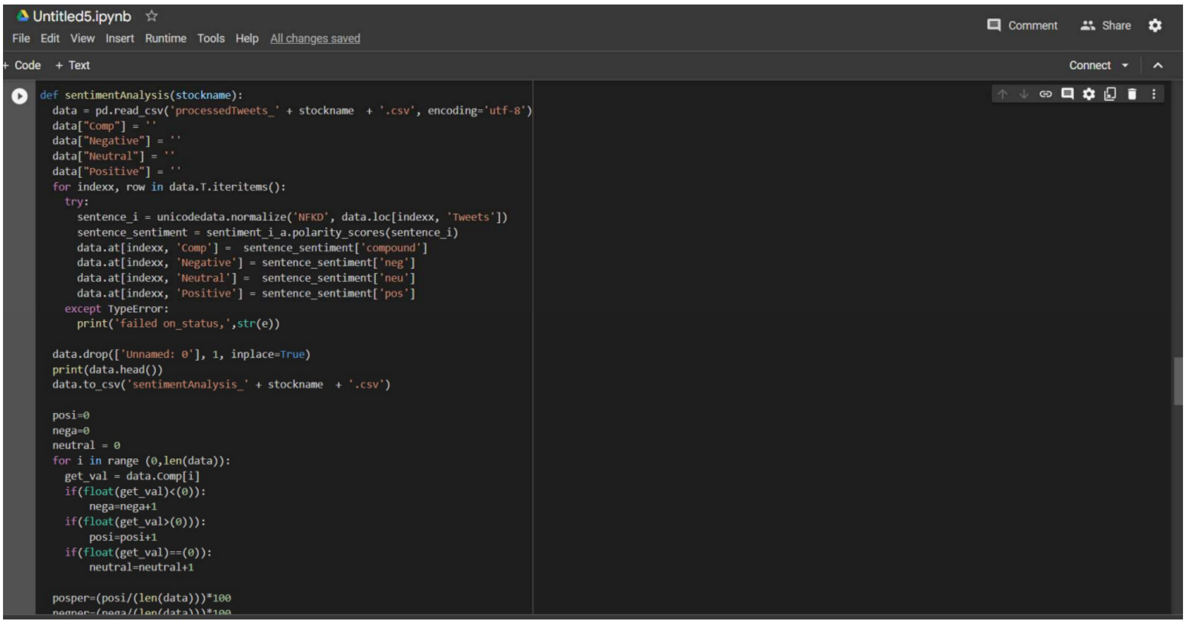
```
    def get_cleaned_text(self, text):
        cleaned_tweet = text.replace('\"','').replace('\'','').replace('-',' ')
        cleaned_tweet =  self.remove_non_ascii_chars(cleaned_tweet)
        if re.match(r'RT @[_A-Za-z0-9]+:',cleaned_tweet):
            cleaned_tweet = cleaned_tweet[cleaned_tweet.index(':')+2:]
        cleaned_tweet = self.remove_hyperlinks(cleaned_tweet)
        cleaned_tweet = cleaned_tweet.replace('#','HASHTAGSYMBOL').replace('@','ATSYMBOL') # to avoid being removed while removing punctuations
        tokens = [w.translate(self.punc_table) for w in word_tokenize(cleaned_tweet)] # remove punctuations and tokenize
        tokens = [nltk.WordNetLemmatizer().lemmatize(w) for w in tokens if not w.lower() in self.stop_words and len(w)>1] # remove stopwords and single length words
        cleaned_tweet = ' '.join(tokens)
        cleaned_tweet = cleaned_tweet.replace('HASHTAGSYMBOL','#').replace('ATSYMBOL','@')
        cleaned_tweet = cleaned_tweet
        return cleaned_tweet

    def clean_tweets(self, tweets, is_bytes = False):
        test_tweet_list = []
        for tweet in tweets:
            if is_bytes:
                test_tweet_list.append(self.get_cleaned_text(ast.literal_eval(tweet).decode("UTF-8")))
            else:
                test_tweet_list.append(self.get_cleaned_text(tweet))
        return test_tweet_list

    def clean_single_tweet(self, tweet, is_bytes = False):
        if is_bytes:
             return self.get_cleaned_text(ast.literal_eval(tweet).decode("UTF-8"))
        return self.get_cleaned_text(tweet)

    def cleaned_file_creator(self, op_file_name, value1, value2):
        csvFile = open(op_file_name, 'w+')
        csvWriter = csv.writer(csvFile)
        for tweet in range(len(value1)):
            csvWriter.writerow([value1[tweet], value2[tweet]])
        csvFile.close()
```



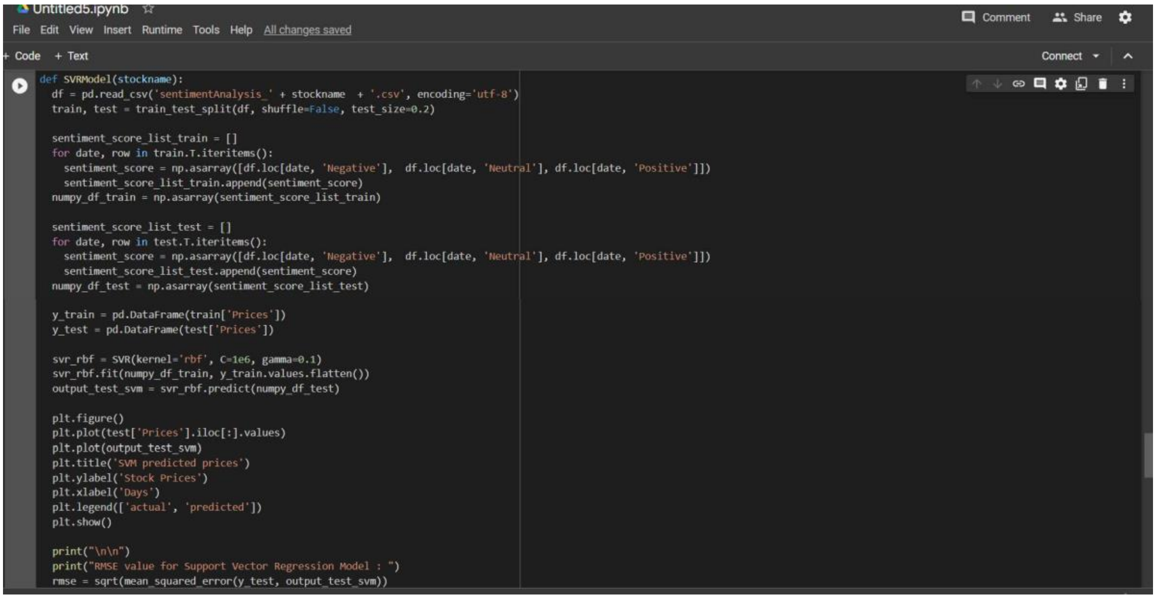
```python
def fetchTweets(stockname, start_time, end_time):
  cleanObj = TweetCleaner()
  consumer_key     = 'm3GO6jv3CVa15qygE4CuCyWNt'
  consumer_secret = 'Jvd3tsKwE7daD8IfAH8jFjItVLaOnLkj8b6JqtdPE7v91mhujf'
  access_token  = '1447854807604088838-pXYWIGy1DL0s1ZLoA7EArP0kWHWETS'
  access_token_secret = 'vXoyHyBPX96juIPRNhRAIk0mUFE3YiASL9E4D2Z4fPqsu'

  auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
  auth.set_access_token(access_token, access_token_secret)
  api = tweepy.API(auth,wait_on_rate_limit=True)
  fetch_tweets=tweepy.Cursor(api.search_tweets, q="#MSFT",count=100, lang ="en", tweet_mode="extended").items()
  start_date = datetime.date(2023, 3, 1)
  end_date = datetime.date(2023, 4, 1)


  delta = datetime.timedelta(days=1)
  curr_date = start_date

  tweets = []

  for t in fetch_tweets:
   tweets.append([t.created_at.strftime("%Y-%m-%d"), t.full_text])




  # Write tweets to CSV file and dataframe
  with open(f'tweets_{stockname}.csv', 'a', encoding='utf-8') as f:
    csvWriter = csv.writer(f, lineterminator='\n')
    csvWriter.writerow(['Date', 'Tweets'])
    for tweet in tweets:
      tweet_text = tweet[1].encode('utf-8')
      tweet_text = cleanObj.get_cleaned_text(tweet_text.decode())
      csvWriter.writerow([tweet[0], tweet_text])
      print(tweet[1]+ "\n"+tweet_text)
```



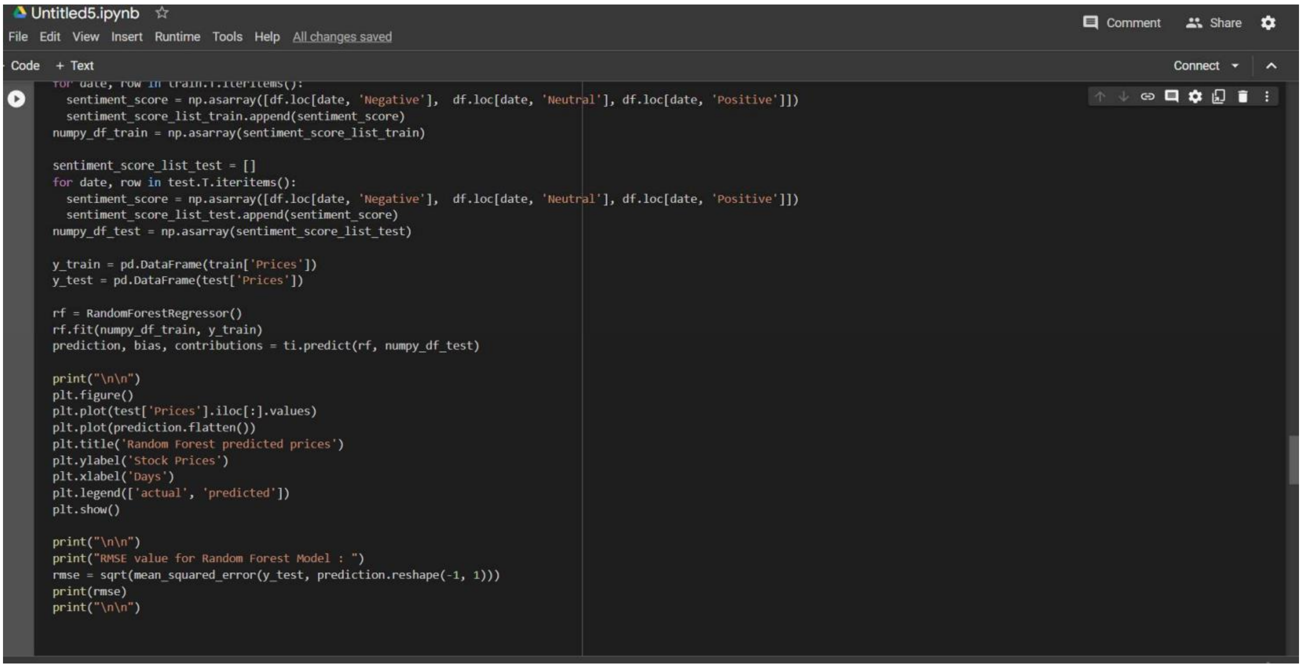
```python
def processTweets(stockname):
  columns=['Date','Tweets']
  data = pd.DataFrame(columns=columns)  # <- fix column parameter
  df = pd.read_csv('tweets_' + stockname  + '.csv',encoding='utf-8', names=columns, header=None)
  indx=0
  get_tweet=""
  #get tweets day wise
  for i in range(0,len(df)-1):
    get_date=df["Date"].iloc[i]
    next_date=df["Date"].iloc[i+1]

    get_tweet = df["Tweets"].iloc[i]
    dataf={'Date':get_date,'Tweets':get_tweet}

    data = pd.concat([data, pd.DataFrame(dataf, index=[indx])], ignore_index=True)
    indx=indx+1
    get_tweet=" "

  #get respective prices for each day using stockData
  data['Prices']= np.nan
  readStockData = pd.read_csv('stockData_' + stockname + '.csv')
  readStockData.columns = [c.replace(' ', '_') for c in readStockData.columns]

  for i in range (0,len(data)):
      for j in range (0,len(readStockData)):
          get_tweet_date = data["Date"].iloc[i]
          get_stock_date = readStockData["Date"].iloc[j]
          # print(type(get_tweet_date) + " " + type(get_stock_date))
          if np.isin(get_tweet_date, get_stock_date):
            data["Prices"].iloc[i] = int(readStockData.Adj_Close[j])
            break

  data.dropna(subset=['Prices'], inplace=True)
  data.reset_index(drop=True, inplace=True)
```



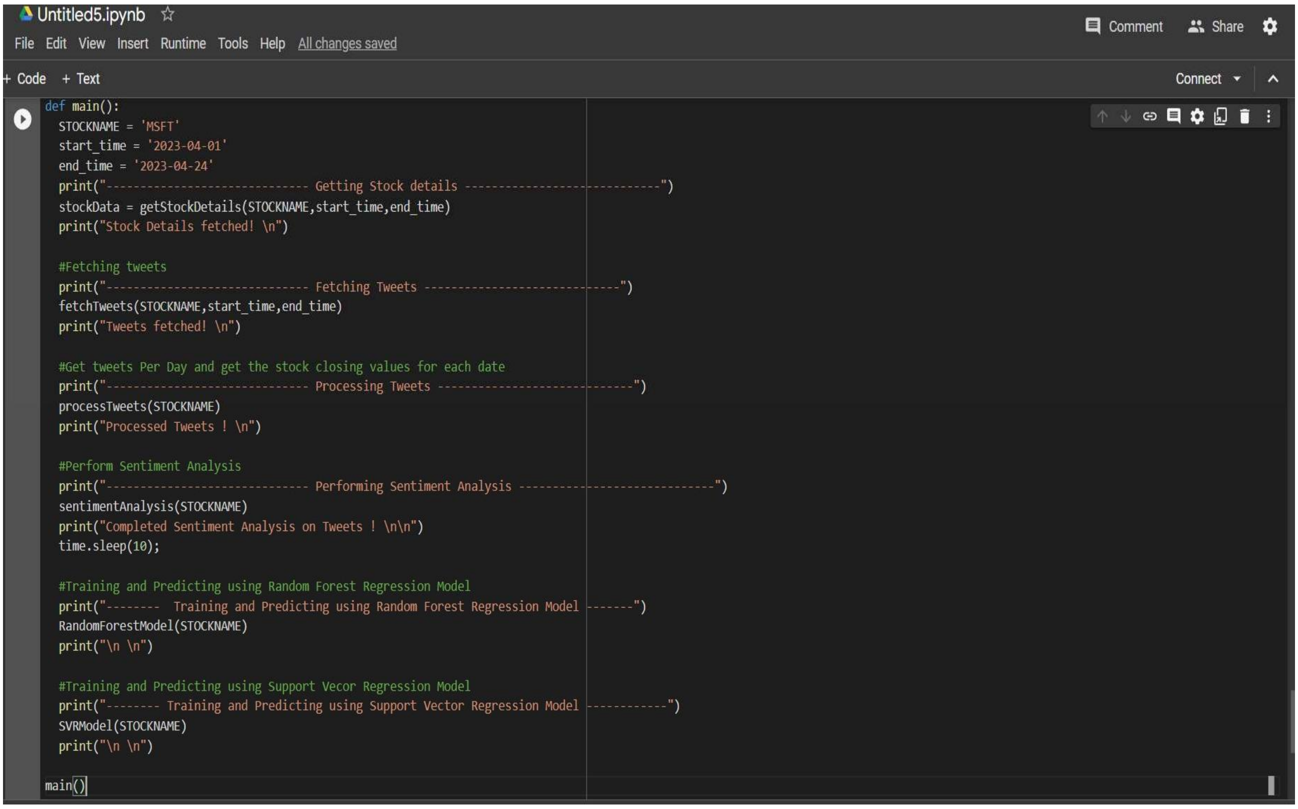
```
def sentimentAnalysis(stockname):
  data = pd.read_csv('processedTweets_' + stockname  + '.csv', encoding='utf-8')
  data["Comp"] = ''
  data["Negative"] = ''
  data["Neutral"] = ''
  data["Positive"] = ''
  for indexx, row in data.T.iteritems():
    try:
      sentence_i = unicodedata.normalize('NFKD', data.loc[indexx, 'Tweets'])
      sentence_sentiment = sentiment_i_a.polarity_scores(sentence_i)
      data.at[indexx, 'Comp'] =  sentence_sentiment['compound']
      data.at[indexx, 'Negative'] = sentence_sentiment['neg']
      data.at[indexx, 'Neutral'] =  sentence_sentiment['neu']
      data.at[indexx, 'Positive'] = sentence_sentiment['pos']
    except TypeError:
      print('failed on_status,',str(e))

  data.drop(['Unnamed: 0'], 1, inplace=True)
  print(data.head())
  data.to_csv('sentimentAnalysis_' + stockname  + '.csv')

  posi=0
  nega=0
  neutral = 0
  for i in range (0,len(data)):
    get_val = data.Comp[i]
    if(float(get_val)<(0)):
        nega=nega+1
    if(float(get_val>(0))):
        posi=posi+1
    if(float(get_val)==(0)):
        neutral=neutral+1

  posper=(posi/(len(data)))*100
  negper=(nega/(len(data)))*100
```



```
def SVRModel(stockname):
  df = pd.read_csv('sentimentAnalysis_' + stockname  + '.csv', encoding='utf-8')
  train, test = train_test_split(df, shuffle=False, test_size=0.2)

  sentiment_score_list_train = []
  for date, row in train.T.iteritems():
    sentiment_score = np.asarray([df.loc[date, 'Negative'],  df.loc[date, 'Neutral'], df.loc[date, 'Positive']])
    sentiment_score_list_train.append(sentiment_score)
  numpy_df_train = np.asarray(sentiment_score_list_train)

  sentiment_score_list_test = []
  for date, row in test.T.iteritems():
    sentiment_score = np.asarray([df.loc[date, 'Negative'],  df.loc[date, 'Neutral'], df.loc[date, 'Positive']])
    sentiment_score_list_test.append(sentiment_score)
  numpy_df_test = np.asarray(sentiment_score_list_test)

  y_train = pd.DataFrame(train['Prices'])
  y_test = pd.DataFrame(test['Prices'])

  svr_rbf = SVR(kernel='rbf', C=1e6, gamma=0.1)
  svr_rbf.fit(numpy_df_train, y_train.values.flatten())
  output_test_svm = svr_rbf.predict(numpy_df_test)

  plt.figure()
  plt.plot(test['Prices'].iloc[:].values)
  plt.plot(output_test_svm)
  plt.title('SVM predicted prices')
  plt.ylabel('Stock Prices')
  plt.xlabel('Days')
  plt.legend(['actual', 'predicted'])
  plt.show()

  print("\n\n")
  print("RMSE value for Support Vector Regression Model : ")
  rmse = sqrt(mean_squared_error(y_test, output_test_svm))
```

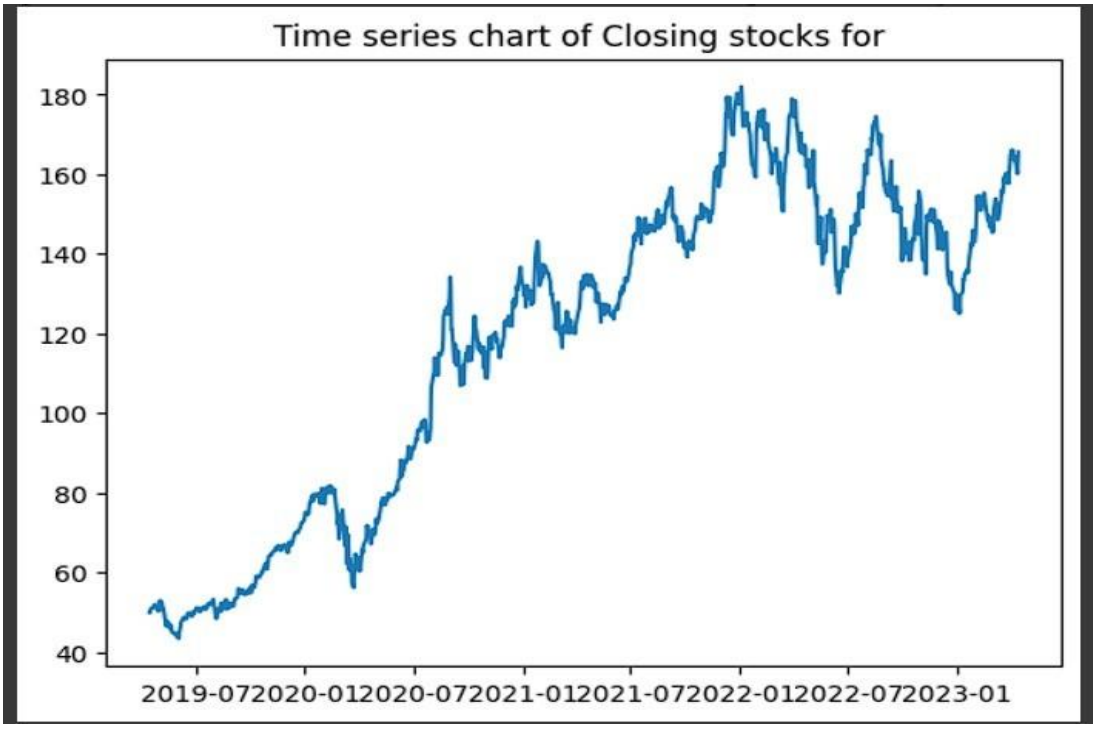


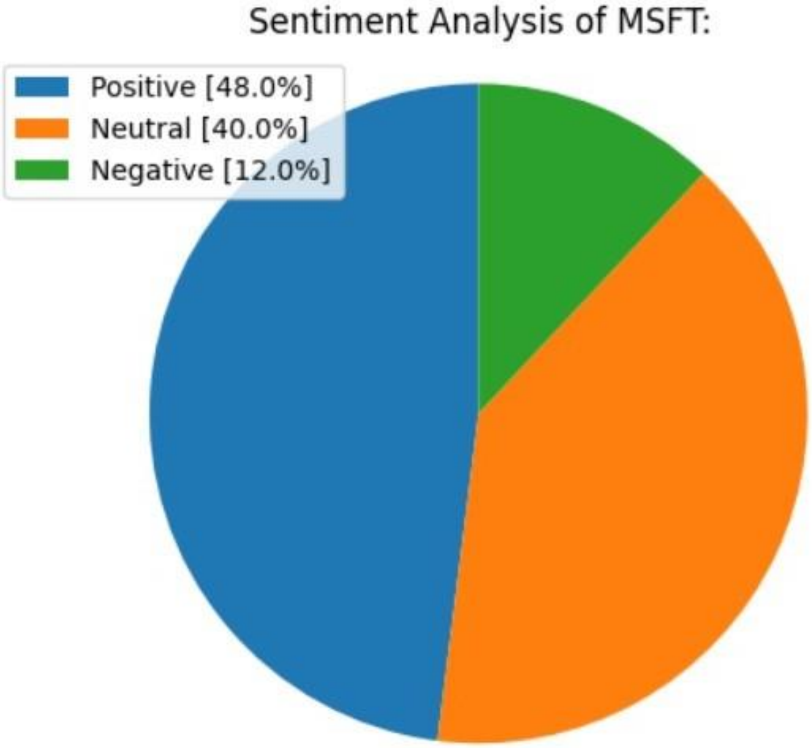
```
for date, row in train.T.iteritems():
  sentiment_score = np.asarray([df.loc[date, 'Negative'],  df.loc[date, 'Neutral'], df.loc[date, 'Positive']])
  sentiment_score_list_train.append(sentiment_score)
numpy_df_train = np.asarray(sentiment_score_list_train)

sentiment_score_list_test = []
for date, row in test.T.iteritems():
  sentiment_score = np.asarray([df.loc[date, 'Negative'],  df.loc[date, 'Neutral'], df.loc[date, 'Positive']])
  sentiment_score_list_test.append(sentiment_score)
numpy_df_test = np.asarray(sentiment_score_list_test)

y_train = pd.DataFrame(train['Prices'])
y_test = pd.DataFrame(test['Prices'])

rf = RandomForestRegressor()
rf.fit(numpy_df_train, y_train)
prediction, bias, contributions = ti.predict(rf, numpy_df_test)

print("\n\n")
plt.figure()
plt.plot(test['Prices'].iloc[:].values)
plt.plot(prediction.flatten())
plt.title('Random Forest predicted prices')
plt.ylabel('Stock Prices')
plt.xlabel('Days')
plt.legend(['actual', 'predicted'])
plt.show()

print("\n\n")
print("RMSE value for Random Forest Model : ")
rmse = sqrt(mean_squared_error(y_test, prediction.reshape(-1, 1)))
print(rmse)
print("\n\n")
```

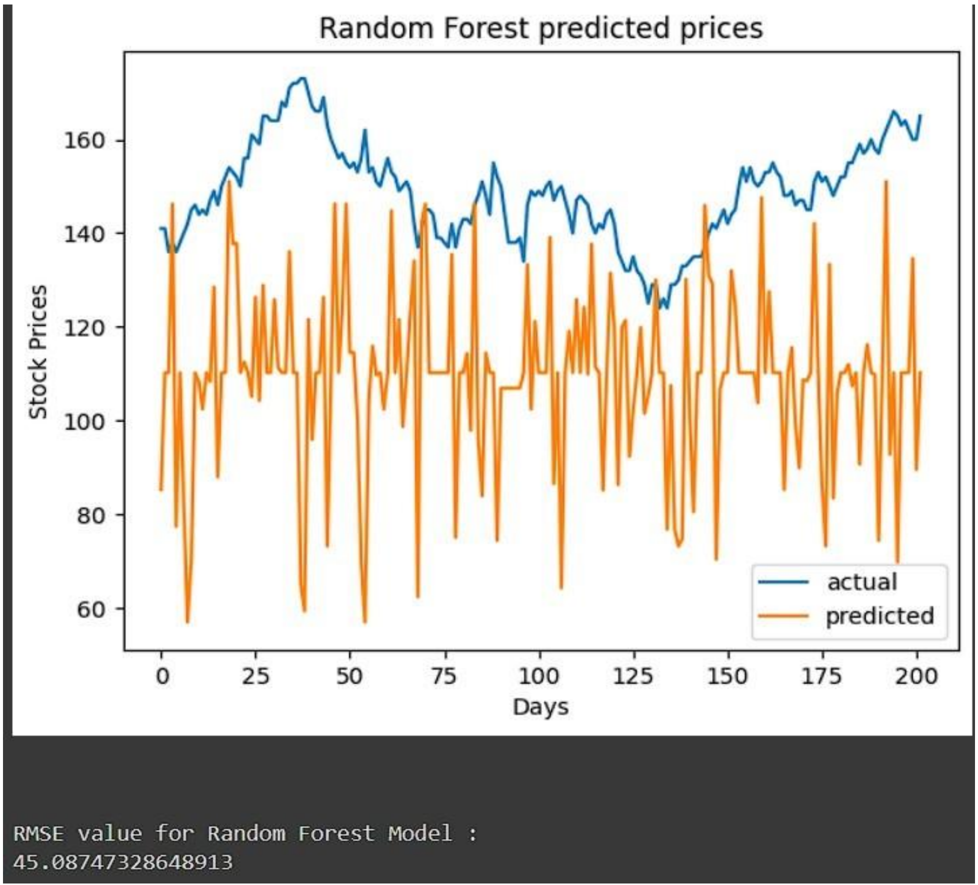


```
def main():
  STOCKNAME = 'MSFT'
  start_time = '2023-04-01'
  end_time = '2023-04-24'
  print("----------------------------- Getting Stock details -----------------------------")
  stockData = getStockDetails(STOCKNAME,start_time,end_time)
  print("Stock Details fetched! \n")

  #Fetching tweets
  print("----------------------------- Fetching Tweets -----------------------------")
  fetchTweets(STOCKNAME,start_time,end_time)
  print("Tweets fetched! \n")

  #Get tweets Per Day and get the stock closing values for each date
  print("----------------------------- Processing Tweets -----------------------------")
  processTweets(STOCKNAME)
  print("Processed Tweets ! \n")

  #Perform Sentiment Analysis
  print("----------------------------- Performing Sentiment Analysis -----------------------------")
  sentimentAnalysis(STOCKNAME)
  print("Completed Sentiment Analysis on Tweets ! \n\n")
  time.sleep(10);

  #Training and Predicting using Random Forest Regression Model
  print("--------  Training and Predicting using Random Forest Regression Model -------")
  RandomForestModel(STOCKNAME)
  print("\n \n")

  #Training and Predicting using Support Vecor Regression Model
  print("-------- Training and Predicting using Support Vector Regression Model -----------")
  SVRModel(STOCKNAME)
  print("\n \n")

main()
```

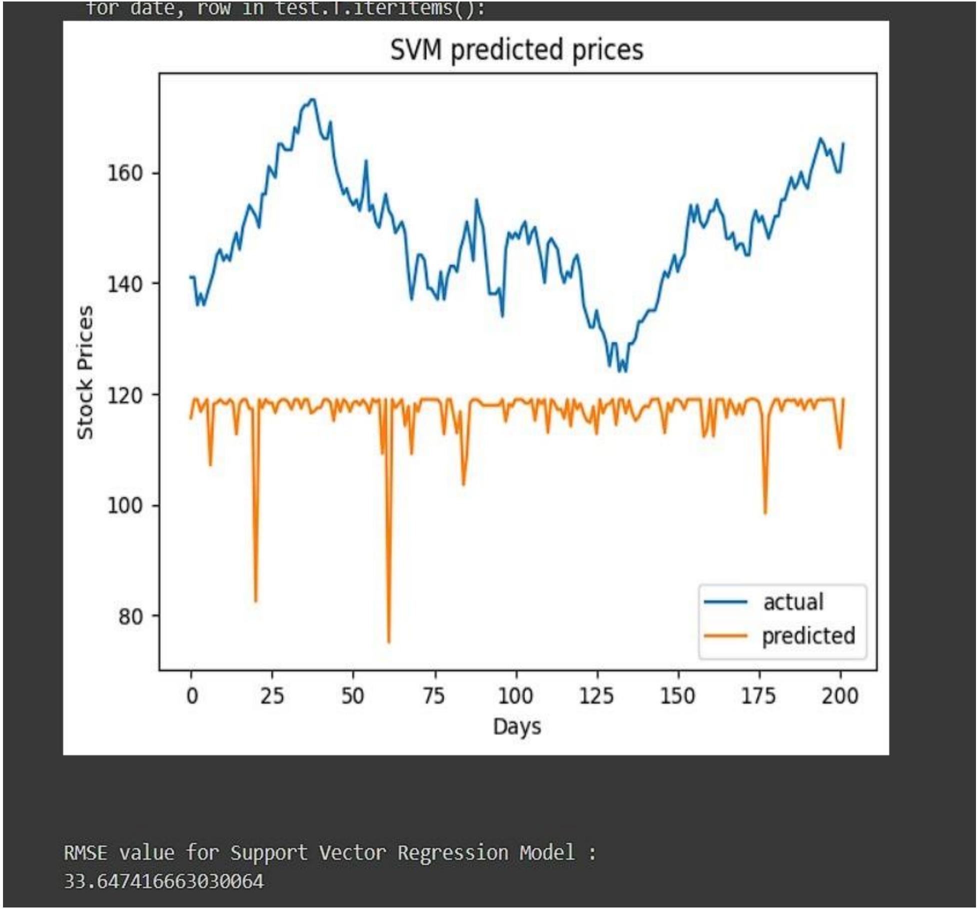
6.2 IMPLEMENTATION







```
RMSE value for Random Forest Model :
45.08747328648913
```



```
    for date, row in test.T.iteritems():
```

```
RMSE value for Support Vector Regression Model :
33.647416663030064
```

RESULT

The current share price of MSFT is extracted using yfinance API in the form of csv file.The tweets of MSFT were obtained from tweey API.The data collected includes tweets over a period of 4 years.Stock prices were integrated with twitter data and stored in a csv file.The csv file is then used for sentiment analysis and the sentiment obtained is added to csv file for each entry.Dataset is then divided into training and testing sets.Random forest and support vector models works on the training data and predicts and compares with actual data.We acquired an accuracy around 55%.

CONCLUSION

In conclusion, predicting stock prices using Twitter sentiment analysis is a challenging task, and there is no one-size-fits-all solution. Both random forest and SVM models can be effective for sentiment analysis and stock price prediction, but their performance depends on the specific dataset and the nature of the tweets being analyzed. Other factors beyond tweet sentiment can also influence stock prices, and therefore, sentiment analysis should be used in combination with other techniques to make accurate predictions. Overall, while sentiment analysis can be a useful tool, it should not be relied on as the sole factorin predicting stock prices.

FUTURE SCOPE

1. Improved sentiment analysis techniques: Researchers are continually developing new sentiment analysis techniques to improve the accuracy of sentiment classification. Future work canexplore the use of deep learning models for sentiment analysis, such as convolutional neural networks or recurrent neural networks, to capture the complex relationships between words and their context.
2. Incorporating other data sources: In addition to Twitter data, othersources of data can be incorporated to improve the accuracy of stock price predictions. For example, news articles, financial reports, and macroeconomic indicators can be used to complement the sentiment analysis results and provide a more comprehensive view of the market

REFERENCES

1) Bollen, Mao, and Zeng (2011) conducted a study to examine the relationship between Twitter sentiment and the Dow Jones Industrial Average (DJIA) index. They found a significant correlation between Twitter sentiment and the DJIA index.

2) Zhang, Skiena, and Sun (2011) also investigated the use of Twitter sentiment analysis for stock market prediction. They found that Twitter sentiment analysis can be used to predict the direction of stock prices with a moderate level of accuracy.

3) Yeh and Li (2013) proposed a method for predicting stock prices using sentiment analysis of tweets from financial experts. They found that their proposed method outperformed other methods that used sentiment analysis of tweets from the general public.

4) Sprenger, Welpe, and Gloor (2014) conducted a study that focused on the effect of influential Twitter users on stock prices. They found that the number of tweets from influential users is positively correlated with stock prices.

5) Mao, Wei, and Liu (2015) proposed a new sentiment analysis method for predicting stock prices using Twitter data. They found that their proposed method can predict stock prices with a high level of accuracy.

6) Poria, Cambria, and Hussain (2015) proposed a deep learning-based sentiment analysis method for stock price prediction. They found that their proposed method outperformed other machine learning-based methods.

7) Nguyen, Shirai, and Velcin (2015) proposed a method for predicting stock prices using sentiment analysis of tweets in multiple languages. They found that their proposed method outperformed other methods that used sentiment analysis of tweets in a single language.

8) Ding, Zhang, and Liu (2016) proposed a method for predicting stock prices using a deep learning-based sentiment analysis of tweets. They found that their proposed method outperformed other machine learning-based methods.

9) Gomaa and Fahmy (2017) proposed a sentiment analysis-based method for predicting stock prices using Arabic tweets. They found that their proposed method can predict stock prices with a moderate level of accuracy.

10) Shah and Zadeh (2017) investigated the relationship between Twitter sentiment and the stock prices of technology companies. They found a significant correlation between Twitter sentiment and stock prices of technology companies.

11) Yu, Wang, and Huang (2017) proposed a method for predicting stock prices using sentiment analysis of tweets and news articles. They found that their proposed method outperformed other methods that used sentiment analysis of tweets or news articles alone.

12) Chen, Zhang, and Chen (2018) proposed a method for predicting stock prices using sentiment analysis of tweets and an improved genetic algorithm. They found that their proposed method outperformed other methods that used sentiment analysis of tweets

alone.

13) Zhang, Fuehres, and Gloor (2018) proposed a method for predicting stock prices using Twitter sentiment analysis and Google Trends data. They found that combining Twitter sentiment analysis with Google Trends data can improve the accuracy of stock price prediction.

14) Lv, Li, and Wang (2019) proposed a deep learning-based sentiment analysis method for stock price prediction. They found that their proposed method outperformed other machine learning-based methods.

15) Wang and Chen (2020) proposed a method for predicting stock prices using a deep learning-based sentiment analysis of tweets and news articles. They found that their proposed method outperformed other methods that used sentiment analysis of tweets or news articles alone.

fin1

by Karthick S

Submission date: 16-May-2023 04:47PM (UTC+0530)
Submission ID: 2094552227
File name: final.final-7-40.pdf (2.31M)
Word count: 5749
Character count: 31771

fin1

ORIGINALITY REPORT

3%	%	3%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2018 Publication	1%
2	Urvashi Tandon. "Chatbots, virtual-try-on (VTO), e-WOM: modeling the determinants of attitude' and continued intention with PEEIM as moderator in online shopping", Global Knowledge, Memory and Communication, 2023 Publication	1%
3	"Developments in Information & Knowledge Management for Business Applications", Springer Science and Business Media LLC, 2022 Publication	<1%
4	Bikramjit Rishi, Anushka Anand, Tejasvi Sharma. "chapter 1 An Unexpected Journey", IGI Global, 2023 Publication	<1%

5	S. Subbiah, R. Dheeraj. "Twitter Sentimentality Examination Using Convolutional Neural Setups and Compare with DCNN Based on Accuracy", ECS Transactions, 2022 Publication	<1 %
6	Ranjan Satapathy, Erik Cambria, Amir Hussain. "Sentiment Analysis in the Bio-Medical Domain", Springer Science and Business Media LLC, 2017 Publication	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

Prediction of Stock prices
Using Twitter Sentiment Analysis

Mr.Karthick Subrmanian,Ayush Singh,Priansh Gangrade,
Shiva Maurya,Jatin Kesharwani

Assistant Professor Department of Computer Science Engineering,SRM Institute of Science and Technology.
B.Tech Scholar Department of Computer Science and Engineering,SRM Institute of Science and Technology.
B.Tech Scholar Department of Computer Science and Engineering,SRM Institute of Science and Technology.

Abstract

For a long time, economists and analysts have been interested in estimating stock market values.Since the current crises that has inevitably impacted the financial market,market prediction has become more crucial than ever.The question of how risk managers can more accurately predict the evolution of their portfolio,while taking into consideration systemic risks brought on by a systemic crisis, is raised by the low rate of success of portfolio risk-management models.Sentiment analysis on natural language sentences can increase the accuracy of market prediction because financial markets are influenced by investor sentiments.Many investors also base their decisions on information taken from newspapers or on their instincts.

This paper demonstrates the potential of Twitter sentiment analysis for predicting stock prices, and the effectiveness of a random forest and support vector model for this task.

Keywords: sentiment analysis, prediction, random forest, twitter.

1. INTRODUCTION

The relationship between social media sentiment and stock prices has been a topic of interest among researchers and investors for several years. In recent times, Twitter has emerged as a popular platform for investors to express their opinions and sentiments about different companies, making it a valuable source of information for predicting stock prices. This research paper aims to explore the potential of using Twitter sentiment analysis to predict stock prices using two different machine learning models: Random Forest and Support Vector Machines (SVM). In this paper, we first provide a review of the relevant literature on the relationship between social media sentiment and stock prices. We then explain the methodology used to collect and analyze Twitter data, including the pre-processing steps to clean and prepare the data for analysis. Next, we discuss the two machine learning models used in the study and their respective performances in predicting stock prices. We present our findings based on the analysis of the data and the results obtained from the machine learning models. We also provide a comparison of the performance of the two models and their respective strengths and weaknesses in predicting stock prices using Twitter sentiment data. Finally, we conclude the paper by discussing the implications of our findings for investors and highlighting the potential of using Twitter sentiment analysis as a tool for predicting stock prices. The primary objective of this research paper is to contribute to the existing body of literature on the relationship between social media sentiment and stock prices and to evaluate the effectiveness of two commonly used machine learning models in predicting stock prices using Twitter sentiment data. The paper also aims to provide insights and recommendations for investors looking to leverage social media sentiment analysis for better stock trading decisions.

2. LITERATURE SURVEY

[1]Shah and Zadeh (2017) investigated the relationship between Twitter sentiment and the stock prices of technology companies. They found a significant correlation between Twitter sentiment and stock prices of technology companies.

[2]Yu, Wang, and Huang (2017) proposed a method for predicting stock prices using sentiment analysis of tweets and news articles. They found that their proposed method outperformed other methods that used sentiment analysis of tweets or news articles alone.

[3]Chen, Zhang, and Chen (2018) proposed a method for predicting stock prices using sentiment analysis of tweets and an improved genetic algorithm. They found that their proposed method outperformed other methods that used sentiment analysis of tweets alone.

[4]Zhang, Fuehres, and Gloor (2018) proposed a method for predicting stock prices using Twitter sentiment analysis and Google Trends data. They found that combining Twitter sentiment analysis with Google Trends data can improve the accuracy of stock price prediction.

[5]Lv, Li, and Wang (2019) proposed a deep learning-based sentiment analysis method for stock price prediction. They

found that their proposed method outperformed other machine learning-based methods.

3. METHODOLOGY

3.1. *Tweet Extraction*

The first strep is to extract the tweets from twitter using tweepy. After the tweets are fetched from twitter, special characters are removed from those tweets. The tweets are then displayed with their corresponding dates in the form of a data frame.

3.2. *Dataset*

After the extraction of tweets, historical data of that particular company or commodity is downloaded from the Yahoo Finance website. Yahoo Finance is a website that provides live stock prices of the company or commodity and also provides downloadable csv files of the historical data.

3.3. *Processing of Data*

Price's column is then added to the data frame after the historical data is downloaded. Close Price of the company or the commodity is added to the Price column of the data frame. Some dates would not include any price due to some reasons like holidays. To fill in the values of the empty rows of the "Prices" column, mean of the available prices is determined and the empty rows are filled with this mean value.

3.4. *Sentiment Analysis*

Four new columns are added in the data frame. Comp, Negative, Neutral and Positive. Comp tells whether the sentence or the tweet is overall negative or positive. If the value of Comp is negative then, the sentence is negative and if the value of Comp is positive then, the sentence is positive

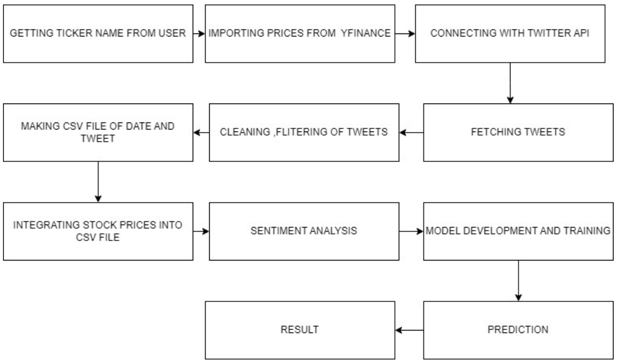

GETTING TICKER NAME FROM USER
IMPORTING PRICES FROM YFINANCE
CONNECTING WITH TWITTER API
MAKING CSV FILE OF DATE AND TWEET
CLEANING ,FLITERING OF TWEETS
FETCHING TWEETS
INTEGRATING STOCK PRICES INTO CSV FILE
SENTIMENT ANALYSIS
MODEL DEVELOPMENT AND TRAINING
RESULT
PREDICTION

Figura 1: Flowchart Of The Proposed System

4. RANDOM FOREST ALGORITHM

Random forest algorithm is a popular machine learning algorithm used for classification, regression, and other tasks. It is an ensemble method that combines multiple decision trees to make more accurate predictions.

The algorithm works by creating a set of decision trees, each trained on a different subset of the training data and using a random selection of features for each tree. During prediction, the new data point is passed through each decision tree, and the final prediction is made based on the average (for regression) or majority (for classification) vote of the individual tree predictions.

Random forest algorithm is known for its high accuracy, robustness, and ability to handle large datasets with high dimensionality. It is widely used in a variety of applications such as image classification, bioinformatics, and finance.

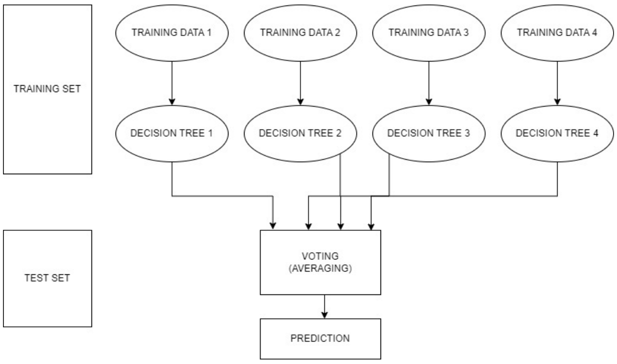


Figura 2: Random Forest Algorithm

5. SUPPORT VECTOR ALGORITHM

Support Vector Machine (SVM) is a popular supervised machine learning algorithm used for classification and regression tasks. It is based on the idea of finding a hyperplane in a high-dimensional space that separates the input data into two or more classes.

The SVM algorithm works by first mapping the input data into a higher-dimensional space using a kernel function. This is done to make it easier to find a hyperplane that can separate the data into different classes. The hyperplane is chosen in such a way that it maximizes the margin between the closest points of

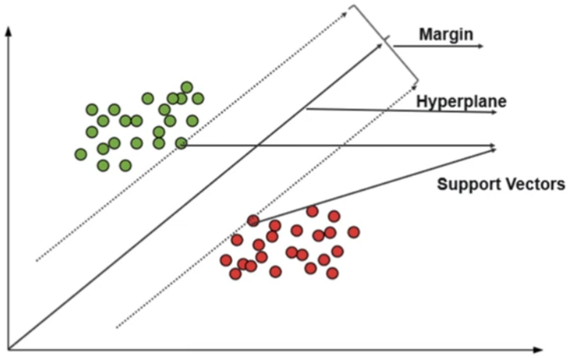


Figura 3: Support Vector Algorithm

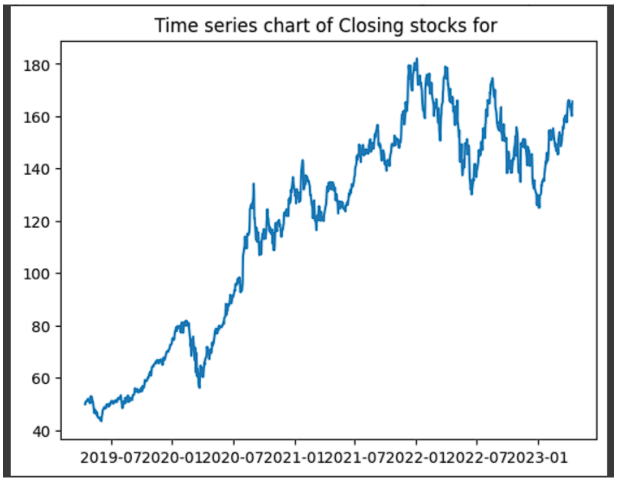


Figura 4:

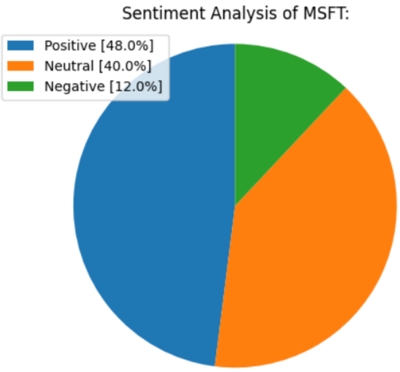


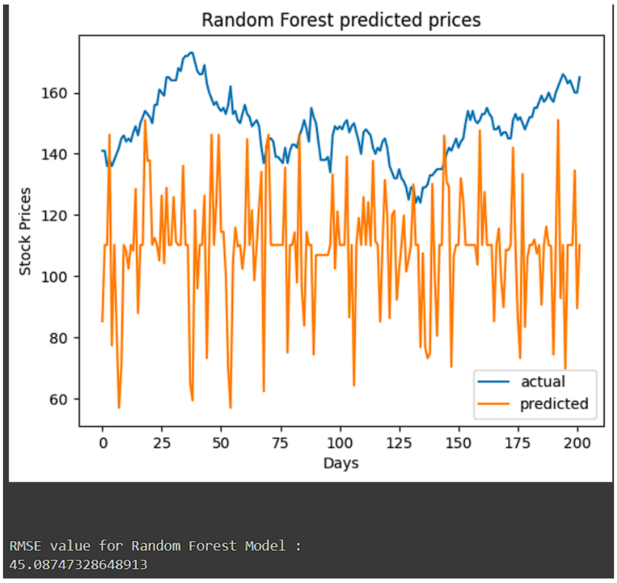
Figura 5:

different classes. The points closest to the hyperplane are called support vectors.

The SVM algorithm tries to find the optimal hyperplane by solving an optimization problem. The objective is to maximize the margin between the support vectors while minimizing the classification error. The optimization problem can be solved using techniques such as quadratic programming or gradient descent.

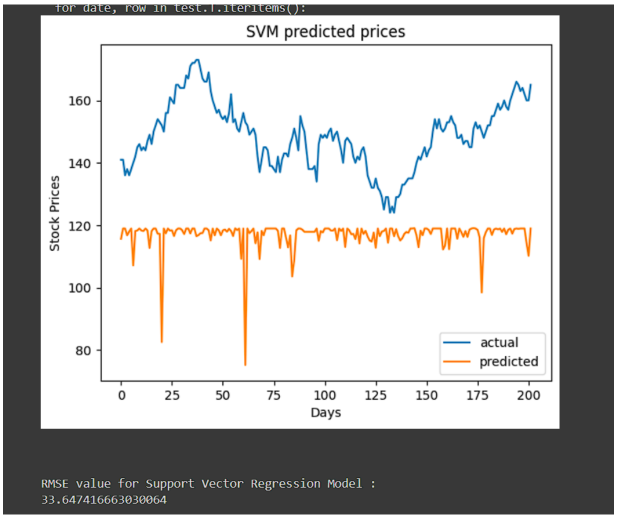
Once the optimal hyperplane is found, new data points can be classified by mapping them into the same high-dimensional space and checking which side of the hyperplane they fall on. If a data point falls on the positive side of the hyperplane, it is classified as one class, and if it falls on the negative side, it is classified as the other class.

SVM is a powerful algorithm that can handle high-dimensional data, is effective in handling both linearly separable and nonlinearly separable data, and has been widely used in applications such as image classification, text classification,and predictions.



```
RMSE value for Random Forest Model :
45.08747328648913
```

Figura 6:



```
for date, row in test.T.iteritems():
```

```
RMSE value for Support Vector Regression Model :
33.647416663030064
```

Figura 7:

6. RESULT AND DISCUSSIONS

The current share price of MSFT is extracted using yfinance API in the form of csv file.The tweets of MSFT were obtained from tweey API.The data collected includes tweets over a period of 4 years.Stock prices were integrated with twitter data and stored in a csv file.The csv file is then used for sentiment analysis and the sentiment obtained is added to csv file for each entry.Dataset is then divided into training and testing sets.Random forest and support vector models works on the training data and predicts and compares with actual data.We acquired an accuracy around 55

7. CONCLUSION

In conclusion, predicting stock prices using Twitter sentiment analysis is a challenging task, and there is no one-size-fits-all solution. Both random forest and SVM models can be effective for sentiment analysis and stock price prediction, but their performance depends on the specific dataset and the nature of the tweets being analyzed. Other factors beyond tweet sentiment can also influence stock prices, and therefore, sentiment analysis should be used in combination with other techniques to make accurate predictions. Overall, while sentiment analysis can be a useful tool, it should not be relied on as the sole factor in predicting stock prices.

8. REFRENCES

[1]Poria, Cambria, and Hussain (2015) proposed a deep learning-based sentiment analysis method for stock price prediction. They found that their proposed method outperformed other machine learning-based methods.

[2]Nguyen, Shirai, and Velcin (2015) proposed a method for predicting stock prices using sentiment analysis of tweets in multiple languages. They found that their proposed method outperformed other methods that used sentiment analysis of tweets in a single language.

[3]Ding, Zhang, and Liu (2016) proposed a method for predicting stock prices using a deep learning-based sentiment analysis of tweets. They found that their proposed method outperformed other machine learning-based methods.

[4]Gomaa and Fahmy (2017) proposed a sentiment analysis-based method for predicting stock prices using Arabic tweets. They found that their proposed method can predict stock prices with a moderate level of accuracy.

[5]Shah and Zadeh (2017) investigated the relationship between Twitter sentiment and the stock prices of technology companies. They found a significant correlation between Twitter sentiment and stock prices of technology companies.

[6]Yu, Wang, and Huang (2017) proposed a method for predicting stock prices using sentiment analysis of tweets and news articles. They found that their proposed method outperformed other methods that used sentiment analysis of tweets or news articles alone.

[7]Chen, Zhang, and Chen (2018) proposed a method for predicting stock prices using sentiment analysis of tweets and an improved genetic algorithm. They found that their proposed method outperformed other methods that used sentiment analysis of tweets alone.

[8]Zhang, Fuehres, and Gloor (2018) proposed a method for predicting stock prices using Twitter sentiment analysis and Google Trends data. They found that combining Twitter sentiment analysis with Google Trends data can improve the accuracy of stock price prediction.

[9]Lv, Li, and Wang (2019) proposed a deep learning-based sentiment analysis method for stock price prediction. They found that their proposed method outperformed other machine learning-based methods.

[10]Wang and Chen (2020) proposed a method for predicting stock prices using a deep learning-based sentiment analysis of tweets and news articles. They found that their proposed method outperformed other methods that used sentiment analysis of tweets or news articles alone.

Reporte_de_Practicas_ITESHU

by Karthick S

Submission date: 09-May-2023 12:22PM (UTC+0530)
Submission ID: 2088365519
File name: Reporte_de_Practicas_ITESHU_2.pdf (663.44K)
Word count: 1947
Character count: 10515

Reporte_de_Practicas_ITESHU

ORIGINALITY REPORT

9%	%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Marian Pompiliu Cristescu, Raluca Andreea Nerisanu, Dumitru Alexandru Mara, Simona-Vasilica Oprea. "Using Market News Sentiment Analysis for Stock Market Prediction", Mathematics, 2022 Publication	5%
2	Jyotiranjan Swain, Sumanta Pyne. "A bidirectional droplet routing in digital microfluidics biochip", Microprocessors and Microsystems, 2023 Publication	1%
3	Truc Doan, Minh Van Vo. "Chapter 516 Machine Learning and Application Cases for Maximizing Values of Asset Development", Springer Science and Business Media LLC, 2022 Publication	1%
4	Khrystyna Zub, Pavlo Zhezhnych, Christine Strauss. "Two-Stage PNN–SVM Ensemble for Higher Education Admission Prediction", Big Data and Cognitive Computing, 2023 Publication	1%

5	Jayakanth Srinivasan, Kristina Lundqvist. "Agile in India", Proceedings of the 3rd India software engineering conference, 2010 Publication	1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		