

Stock Market Analysis Using Twitter Sentiment

by

Examination Roll:
Md. Soliman Hossain (172485)
Munira Ferdous (172514)
Adiba Masud (172472)

A Project Report submitted to the
Institute of Information Technology
in partial fulfillment of the requirements for the degree of
Bachelor of Science in Information Technology

Supervisor: Manan Binte Taj Noor,
Assistant Professor



Institute of Information Technology
Jahangirnagar University
Savar, Dhaka-1342
March, 2022

DECLARATION

We hereby declare that this thesis is based on the results found by ourselves. Materials of work found by other researcher are mentioned by reference. This thesis, neither in whole nor in part, has been previously submitted for any degree.

Md. Soliman Hossain
Roll:172485

Adiba Masud
Roll:172472

Munira Ferdous
Roll:172514

CERTIFICATE

This is to certify that the thesis entitled **Stock Market Analysis Using Twitter Sentiment** has been prepared and submitted by **Munira Ferdous, Md. Soliman Hossain** and **Adiba Masud** in partial fulfilment of the requirement for the degree of Bachelor of Science (honors) in Information Technology on March, 2022.

Manan Binte Taj Noor
Supervisor

Accepted and approved in partial fulfilment of the requirement for the degree Bachelor of Science (honors) in Information Technology.

Prof. Dr. Md. Abu
Yousuf
Chairman

Dr. Shahidul Islam
Member

Dr. Rashed Mazumder
Member

Prof. Dr. Md. Hasanul
Kabir
Member (External)

ABSTRACT

Recently happening the epidemic which have impacted the importance inquiring ability of our daily life styles. Sentiments have clothed to be a huge impact at the motion of the inventory change and pandemic has satisfactory introduced greater steam. This study with the limelight at the recent pandemic is a try to analyze the magnificence accuracy of determined-on algorithms for sentiment evaluation and prediction for the inventory costs. We've attempted to take a look at proposed framework for sentiment analysis and prediction for the ideas to discover the correlation among people and forum sentiment. We use twitter records to anticipate public temper and use the anticipated mood and previous days. We have considerably utilized DJIA values for a certain time period and also tweet values for understanding people's sentiment on which we are looking ahead to the inventory market actions. After this, checking highlights the accurate process or methods or algorithms based mostly on processed outcomes. Then the ones algorithms may be enough to give input for building active prediction systems such we want.

Keywords: Stock Market, Epidemic, Sentiment Analysis, Twitter Data.

LIST OF ABBREVIATIONS

API	Application Programming Interface
LR	Linear Regression
LSTM	Long Short-term Memory
DJIA	Dow Jones Industrial Average
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
SOFNN	Self-Organizing Fuzzy Neural Networks

LIST OF NOTATIONS

β	Define Beta
ϵ	Define Epsilon
ω	Define Omega
θ	Define Theta

LIST OF FIGURES

Figure

3.1	System Model Block Diagram	14
3.2	System Architecture	16
3.3	Dataset Pre-processing	17
3.4	Collected DJIA Values	18
3.5	Twitter Users Growth	19
3.6	Textblob Working Process	22
3.7	Sentiment Analysis Process	23
3.8	Pearson's correlation coefficient	25
3.9	Linear Regression Workflow	26
4.1	DJIA Values Stock Price Average	27
4.2	Some Extracted Tweet	28
4.3	Sample Cleaned Tweets	29
4.4	Subjectivity and Polarity	30
4.5	Sentiment Analyzer values	31
4.6	Sentiments of tweets	31
4.7	Predicted Percentage	32
4.8	Actual vs Predicted	33
4.9	Actual values vs Predicted Values	34

TABLE OF CONTENTS

DECLARATION	ii
CERTIFICATE	iii
ABSTRACT	iv
LIST OF ABBREVIATIONS	v
LIST OF NOTATIONS	vi
LIST OF FIGURES	vii
CHAPTER	
I. Introduction	1
1.1 Background	1
1.2 Overview	3
1.3 Problem Statement	4
1.4 Motivation	4
1.5 Objective	4
1.6 Research Outline	5
II. Literature Review	6
2.1 Introduction	6
2.2 Research Gap	11
III. System Model	13
3.1 Overview	13
3.2 Methodology	14
3.3 System Design	15
3.4 Data Set and Preprocessing	15
3.4.1 DJIA Value	17

3.4.2	Twitter Data	19
3.5	Sentiment Analysis	21
3.6	Hypothesis and Final Co-relation	24
3.7	Linear Regression (Proposed Algorithm)	26
IV.	Result Analysis	27
4.1	Comparison with Other Works	34
V.	Conclusion and Future scope	36
	References	38

CHAPTER I

Introduction

1.1 Background

Almost two and a half trillion stock investors within the world per 2021 recent statistics are contributing to global economics through the exchange. The overall global stock market trading turnover is around \$95 trillion. Particularly, retail investors have made a buy or sell investment decisions depends on the world situation. Within the whole world, there are plenty of taking resolutions for an investment to involve an outsized volume of cash are produced. Spending a lot of time on uncovering funding lucky chances by retail shareholders has become a day-to-day difficulty. On the other hand, other investors are looking for white-collar commercial advisory services, the prices are sky high only for retail shareholders. For this reason, they are not being able to take that service easily as others and have to make all types decision depending on their own market analysis. So, this type of doing own analysis is becoming very tough in this modern time when there are a lot of technological opportunity.

Humans are being sophistical in their behavior at some certain events or matters. While not being valued, task based different thinking, choices get biased by psychological feature biases or personal emotions, resulting in gratuitous losses. The analysis of the company's shares is currently obsessed with most on completely different social media sites. Among different social media sites that has most significant among the economic and share market empire is Twitter. Almost 500 million tweets are sent by active users on a usual. To express their various emotions through these tweets, which are being translated into helpful info. At now, such an improbable quantity of social media knowledge can't be entirely accumulated by investors only being alone. It is a just about not possible task for them to make it on their own. Therefore, a processed

analysis system is vital for investors, because this system can mechanically appraise stock trends victimization such giant amounts of knowledge in our worked database.[1]

Without access to quantitative and information-pushed models on people's sentiment analysis, one apparent technique retail buyers could use to wager the market is through simple signs, as an example, easy regression and exponential shifting average. Again, there are huge up-downs in the world stock market for last years, due to coronavirus and for it's post covid impact. Like an example, the stock market came crashing down suddenly because of sudden attack of coronavirus and also the finance instability. This crash because caused a huge fall down to the market and after two months again tried to enter at the stock markets, which continued through at the end of the same year. Another noticed matter is when in 2019 the leading company was Microsoft, but in the middle of year 2020, Arabian oil company took the leading which was worth around \$1.36 trillion.

To get the analysis of the above information and also for closing prediction of stock prices there have been a lot of experiments caused in these recent years. Supervised learning and unsupervised learning classifications of machine learning models are used consequently to predict the stock prices with the most accurate level. Among those several techniques like the Decision Tree method, Support Vector Machine, KNN, Random Forest method, Naive Bayes method, Logistic Regression are vastly used to apply machine learning model for better prediction. Without these techniques, R programming is also used to plot the data of analyzing and also future prediction.

Again, retail investors will be able to predict the stock exchange by drawing a regression line that connects the utmost or minimum of possible prices. A regression system is used to guess and predict the current market trend which is inspired by the adding modern machine learning algorithms for research, these algorithms might serve as implicit tools to seek out actual patterns within the trend of stock prices, this information may be useful to supply redundant perceptivity for retail investors when making different investment planning.

Twitter sentiment analysis is an imperial of data and can hand over intuition that can intimate positive or negative discern on stocks and direction. There has an enough number of research on sentiment analysis on different topics, just for movie reviews and Twitter post in random or defined days, months. In this work we use

sentiment analysis on twitter random posts. Saving Twitter Sentiment Analysis for implementation by applying twitter API kept by twitter. This method is extirpating tweets from twitter in given below :

- Firstly, import essential packages and libraries.
- Determine criterion which build up relationship with twitter API.
- By removing some words for cleaning the tweets.
- Tokenize every phrase withinside the dataset and keep it into the dataset
- For every word to compare with positive or negative.
- Finally, Find the sharpness of positive, negative and neutral tweets.

1.2 Overview

The study is going to connectivity bounded by the recent timeline from 2020 and the stock market of research. In this study, we use sentiment analysis for prediction stock market ups-downs and also use tweeter data that is taken from Twitter. We propose here a completely unique topic modeling throughout sentiment scoring methods. Our proposed model has been designed to function as a number one indicator to serve in recession prediction models. The work is predicated on our hypothesis that includes such a sort of sentiment indicator, with unstructured data aggregated from Twitter trending posts, shall be significant in improving prediction capabilities. As such, derivations made up of unstructured data collected and collated from Twitter posts using Twitter API to supply a measure of the polarity of the type of data that both producers and consumers encounter. Through our research, we demonstrate text sentiment analyses tools while, at an equivalent time, presenting our design for a replacement time-series measurement of economic sentiment. Our data is extracted from yahoo finance DJIA's historical data between March 1, 2020, till now. We undertake comparisons for estimating predictive accuracy of several sentiment analysis models using Twitter data that are posted either as being positive or negative by humans which can almost fall down or up the Dow Jones Industrial Average (DJIA).

1.3 Problem Statement

- Some researchers explored correlation between sentiments of twitter and stock indices but they were unable to meet the main goal.
- For the social network graph, a data visualization tool such as NodeXL was utilized to visualize the result of user's opinion.
- In previous research, presented a Machine Learning (ML) approach that taught using publicly available stock data to build intelligence. In this respect, the study employed a machine learning technique known as Support Vector Machine (SVM) to predict stock prices for large and small capitalizations, as well as in three separate markets, using daily and up-to-the-minute prices.
- In foregoing paper, presented a hybrid technique that combined an LSTM network with a genetic algorithm (GA), they used daily Korea Stock Price Index (KOSPI) data to test the proposed hybrid strategy.
- To establish the relationship between "public sentiment" and "market sentiment" "using sentiment analysis and machine learning concepts and validate their findings, present a new cross validation method for financial data and use SOFNN to achieve 75.56 percent accuracy on Twitter feeds and DJIA values.

1.4 Motivation

The final aim of our work is attending to distribute buyers in the act of any type of important event stab mechanism that makes use to assist the rapid-changing by voyaging stock market of using machine learning. The study pursuits to socialize and open up today's system by getting to know technologies for retail traders. Apart from the models which tried to decrease, exceptional finance-particular rankings are added to segregate and figure out the overall achievement of conflicting systems, especially version accuracy rating, version trend score, and stock purchase or promote rating. The ratings also are constructed to bring important and consequentially words to assist buyers to apprehend inventory and make funding decisions.

1.5 Objective

This proposed system will serve the following objectives :

- To propose a novel topic modeling accompanied by sentiment scoring methods, designed to function as a leading indicator, and to serve in recession prediction models.
- To construct a model having reduced complexity and higher accuracy.
- To create this model, an appropriate algorithm text blob was used to predict sentiment.
- To assemble as an additional perceptible contraption for investors from a unique mindset with the assist of the era they'll be looking for better statistics in peer the market.
- A large mess that can give over data with profitable predictions in contrast to public opinion to a suitable circumstance, despite the fact that each tweet is irrelevant to the point of a section.

1.6 Research Outline

Our study is divided into four sections with different sub-sections and also demonstrated the required figures into it. In chapter II, we have discussed related works in this area and which and how the model was used in those studies with proper explanation. Also, we tried to cover the gap between those papers which might be done better for doing the appropriate research. After that, we have briefed our own system model with the required algorithms and figures to better understand and also got the result after applying the machine learning technique in chapter III. Last but not the least, we concluded with our future work by discussing its possible scope and betterment in chapter IV.

CHAPTER II

Literature Review

2.1 Introduction

For security market analysis, there's a requirement to develop new tools that in combination with ancient prediction models can tune the predictions by taking into consideration factors that don't directly have their origin within the company itself, however the final public's perception of the market and opinions of the studied stock especially. a trial to extract additional information from a dataset of tweets was conducted and therefore the chance to acquire higher stock value prediction with a restricted-sized knowledge. The aim of this work is to push a machine learning approach with deep learning to unravel this downside. Typically, once a machine learning model isn't correct enough, it may be solved by coaching it with a bigger dataset. Earlier analysis has used solely volume and follower count. This work extends their findings to check the performance of stock predictions with a machine learning model, using tweet attributes. For prognostication stock market expression and costs progression which have been learned respective vestibule in literature. We studied some works which are focused on developing the process of prediction situated on sentiment analysis of random news or random tweets post both worked on stock costs. Another target on stock market price prediction using several time frames. Furthermore, particular research methods demonstrated that works were a vigorous correlation interpolated stock market prices changing with random twitter posts and both researchers used different formulas or different algorithms. Microblogging has been spreading online and lots of researchers have shown their interest in this means of human reactions for almost 10 years. Due to its capability to transmit concepts across individuals, a quest identifies it as on-line spoken disapproval. Twitter has fairly often been thought about as the foremost simple alternative by researchers for sentiment analysis and opinion mining on microblogging knowledge. Indeed, it pro-

vides a large volume of narrow-minded knowledge on a really broad variety of subjects and encompasses a free API for creep. Tweets and users.

Siersdorfer et al. [2] proposed that a small number of researchers who have percolated sentiment analysis of social networks such as Twitter and YouTube [3], [2], [4]. In dispersion through the above mentioned, the elbow grease which is uttermost in conjunction with affiliate. The researchers who are hypothesized innumerable 6 million comments, which are self-possessed from 67,000 YouTube videos to pick-out the consensus in the thick of comments, views, comment ratings and topic categories. The authors panoply betrothal repercussions in prognosticate the comment appraisal of new modish comments by physique divination models practicing by the time mentioned scaled comments. Nonetheless, the meticulousness of sentiment apportionment inclines condensed of the accuracy of run-of-the-mill topic-based text categorization which mileage such machine learning techniques. Another Pronounced work in sentiment analysis is by Bollen et al. The authors used two sentiment hunting paraphernalia flawlessly crystal-ball divine the circadian reconstruction to the closing scruples of the Dow Jones Industrial Average (DJIA) and Index the Twitter feeds of users. The authors calculate an accuracy of 86.7% and a shrinkage of more than 6% in the mean average percentage error. Our work bears no resemblance intrinsically from the top of the whole shebang.

Nir. B. et. al [5] proposed sentiment analysis that clenched weighing text of frame to take sentiment scores for articles. They primarily deleted pull up words before making up data structure from body text of body. Then they built a verdict TextBlob library. Textblob is made on origin in python of the NLTK package⁷ and grasped for an array of fiction articles to use as a probable tool. Working a lexicon basically foregoing dictionary of contentions-based dictionary supported path at any time, confidential based their polarity values, to count the polarity score of word of a section stunt word Textblob, their python tool obtained proscribe away of a bit of scripting to benefit a packet of texts way to arrive to the sentiment polarity scores. Those scores were to reach the variance of -1 to 1, where -1 defines an intensely negative article although 1 defines a positive article. The scores which were assembled and gripped on in a strikingly learning of pandas frame well-organized with the highlight line of article, the text calculate and hence the parts of the above mentioned works which are permeated to subsequently of large amount of whole, they collated the previous polarity price learning frames for yearly block of news works.

An. Mt. and Al. Mi. [6] et. al wrote that Sentiment analysis was a significant portion solution. In their article, they used four mood parts, such as Calm, Happy, Alert, and Kind. Then they demonstrated a few typical tools such as Opinion Finder, Sent Wordnet etc. searched them deficient, incomplete and hence determined their own analytical code to improve.

Ar. K., Jo. Za. and San. Krs. [7] et. Al proposed that an opinion is well-advised such as a packet of free words. The positive and negative opinions in the suite dataset which are reserved in two partition lexicons, that they excerpt to such as positive lexicon (positive opinions) and negative lexicon (negative opinions). For every opinion, every text is counted the digit of times the word by calculating and the polarity or sentiment arrives in the positive and negative score. Every text, the polarity that is the positive number of times the word issues on a positive score divided by the whole number times that exposes both positive and negative scores.

Ai .V and Dr. Vani Priya [8] et. al proposed that Twitter is passed down on the point of a actual time which confesses people to exchange their surveys, depicts their inquiring mind, regards or inference just as to how they find around any item and imply with brief messages or sectional realm of text tell that constitute of sacralizing service pulpit of 140 characters or less than. Twitter is an excellent data source to be considered for sentiment analysis for predicting stock values. So, brings to tweets from Twitter API, first needs to register an App over their twitter account and specify an application. To follow the steps and then follow to get to the link (<https://developer.twitter.com/en/apps>) and achieve the APIs. Dev Shah et al. [9] et. al had learned about twitter data from twitter newsfeed and applied sentiment analysis on the news. This paper they had discovered polarity for the pharma sector and on the whole their main focus was on the stock's cost ups and downs that was situated on the polarity bookwork.

Another paper written by Du Peng [10] et. al mostly learned the market vaporousness and also created the sentiments of people and to identify the relation.

Vai. Gururaj, Shr. V R and Dr. As. K. [11] et al. proposed that here one in all ways used the $p = mr + q$ equation to suit a flat-out line for structure: a sketch such as row occurs of points of the given dataset via the most number. After that,

a right away line per the dots such a group of the intervening within every drop and consequently the series is least, using mathematical terms, to plot rate of the dataset on an epistle. To call the hypothesis is able to predict the y value for any given x and this prediction technique is mentioned in rectilinear regression and hence the process conducted using the youngest volume squares method. The method is extensively owned by statisticians, availed simultaneously the initial notion of ML. The hypothesis function of rectilinear regression formation in total,

$$p = t\theta(r) = \theta_0 + \theta_1 r \quad (2.1)$$

comment that is mostly similar to the line equation. θ_0 and θ_1 is conferred to θ $t(x)$ to inspire the inventory output p . The validity of their hypothesis is a cost function applying of magnitude, that is, a mean of every result with inputs from r is parallel to exceptional output p is of the hypothesis

Yahya Eru Cakra and Bayu Distiawan Trisedya [12] et al. proposed that Linear regression is that model which is used regression method to be used for classifying numerical class that is created linear function by enumerating weight values (ω) for each feature (β). The function can be written as be subsequent to,

$$x = \beta_0\omega_0 + \beta_1\omega_1 + \dots + \beta_n\omega_n \quad (2.2)$$

X are regression values for example to data. Plethora methods to estimate a linear regression model which cut out of ultimate data has a normal distribution in its enduring. At the edge of, it can be in conjunction with appraised by stretching toward at the coefficient of single-mindedness value (R^2). R^2 which is the square of the coefficient of correlation (R) that is between approximated values and real values. R^2 magnitude from 0 to 1. The more R^2 value close to 1 is the over and above data which is adapted. The coefficient of indomitability which is the lagniappe of the full amount contradicts distinction in the tied to apron strings variable which can be calculated by departure from the norm in the independent variable(s). After that R^2 is + 1, there endures a put finishing touch on linear interrelation between p and q , i.e 100% of the variation in q is elucidated by variation in p . Meanwhile this is $0 < R^2 < 1$, that is an insubstantial linear consanguinity in the thick of p and q .

U. Pr. Gur. and S. K. et. al. [13] wrote that they used the Dow Jones index flush smallest dot of the year during 21% to date or 39%. They cut down 3 months considering the Dow Jones shocked slag sole to return to its previous vertex. David Valle-Cruz, Vanessa Fernandez-Cortez, Asdrúbal López-Chau and Rodrigo Sandoval-Almazán et. al. [14] used Investor sentiment could potentially predict the direction of indexes a few days ahead of time. During the COVID-19 epidemic, markets took 0 to 10 days to react to information published and distributed on Twitter, according to this analysis. This period ranged from 0 to 15 days during the H1N1 pandemic. They discovered correlations not only in the positive but also in the negative shift values (from 11 to 1). This suggests that the stock market's performance influences Twitter users' reactions. This took 1 to 11 days in the case of H1N1, and 1 to 6 days in the case of COVID-19.

Chetan Gondaliya, Ajay Patel² and Tirthank Shah^{the} [15] et. al. used Logistic Regression and Support Vector Machine algorithms which have produced better results, making these two algorithms superior in sentiment prediction when using the Bag-of-words technique. The study makes a substantial contribution to the identification of superior algorithms for sentiment prediction in terms of accuracy and gives a strong platform to evaluate six selected algorithms with special reference to Indian stock market news and covers the most recent time period of Covid-19 pandemic impact, as natural language processing becomes a more powerful tool for interpreting text messages. In an additional study performed by Wang, X., & Luo [16] et. al. employing sentiment analysis to forecast movie performance based on data from social networking sites. They gathered sentiments from Twitter and YouTube, among other social media platforms. The K-means clustering technique was used to predict the outcome of the film. In further study by Aditya Bhardwaja, Yogendra Narayanb, Vanrajc, Pawana and Maitreyee Duttaa et. al. [17] employed the Python scripting language, which has a quick execution environment, to assist investors in predicting where their money should be put in the stock market. They also did not reach their desired goal.

Papadamou, Stephanos and Fassas, Athanasios and Kenourgios, Dimitris and Dimitriou et. al. [18] wrote that the risk-aversion channel of pandemic spread in the stock market, as well as the attention-induced price pressure hypothesis, are both supported by our findings. As a result, our findings add to earlier research that shows how investor attention in a "google" or "internet" oriented economy affects implied

volatilities in stock markets. These findings reveal an investor sentiment channel that grows as a result of behavioral biases during pandemic crisis periods, providing useful information to investors and policymakers. Understanding the links between investors' decisions on a pandemic and asset price volatility is crucial for creating and executing market and economic policy actions.

2.2 Research Gap

V Kranthi Sai Reddy et. al.[19] proposed SVM algorithm works on a big dataset value that is collected from different worldwide financial markets, and it proposes the use of data collected from multiple global financial markets with machine learning algorithms in order to anticipate stock index movements. The main problem with the work is that overfitting is not an issue while using SVM. Various machine learning-based models are currently being presented for predicting the daily movement of market equities, with numerical results indicating a high level of efficiency. Linear regression, on the other hand, is a model that is used to enhance a well-trained predictor. Compared to the chosen benchmarks, the model yields a bigger profit. Thien Hai Nguyen and Kiyoaki Shirai et. al.[20] suggested a TSLDA-based approach for predicting stock price movement based on social media sentiments. Although the proposed method's accuracy of 56 percent is not particularly great, the results can be satisfactory, as evidenced by prior works. However, one disadvantage of TSLDA is that it requires you to specify the number of subjects and emotion ahead of time. To overcome this, a model known as linear regression is a non-parametric topic model that estimates the number of topics present in the data.

P. K. Singh, A. Sachdeva, D. Mahajan, N. Pande, and A. Sharma et. al.[21] wrote that Sentiment analysis was used in a recent study to filter out unnecessary reviews on popular e-commerce platforms. MongoDB was used to store results for unstructured data in the backend of their study. This is an old method of gathering data and performing sentiment analysis. on the other hand, using sentiment analysis on Twitter trending topics, has proven to be quite effective. Aishwarya .V and Dr. Vani Priya et. al. [22] experimented that There are a few things to consider: Sentiment Analysis focuses on using context, tone, and other criteria to forecast stock values and people's opinions in tweets. They developed TextBlob, a model that processes live tweets and categorizes them as good, negative, or neutral, based on Twitter data. This trained model then forecasts stock market rates. They did note, however, that the ratio of negative tweets is substantially larger, and that the results would

be less accurate, hinting that the emotion of the people they analyzed isn't totally accurate. Sandipan Biswas, Prasenjit Das, Rajesh Bose and Sandip Roy et. al.[23] explained the impacts of a pandemic were studied using real-life examples of the effects of coronavirus infections over the world. They studied a variety of news stories with numerical values in order to better comprehend stock market trends based on previous patterns. They also took into account the ideas and opinions expressed by share reviewers. The opinions of reviewers have an impact on traders who invest in the stock market. Their work has some flaws, such as the use of aggregated share prices and the lack of daily news and articles gathered from different financial and stock market websites. Our work differs from the previous work in a few ways; now we may give a brief description of it. We begin by downloading stock market data in a csv file from Yahoo Finance and organizing it intelligently. Then it may be used to generate price average orders using DJIA values, sentiment analysis in Twitter random posts piece by piece, and convincingly forging linear regression for prediction. As a result, we use textblob to collect random tweets for sentiment analysis. We noticed that a big amount of text-file data had been circulating and had shown to be beneficial to researchers. Then, using the Twitter API, acquire random twitter data as well as random twitter postings, clean twitter datasets, then perform TextBlob sentiment analysis to these cleaned datasets. TextBlob assesses and ranks positive, negative, and neutral text. Then calculating the perfect percent of sentiment. Prediction using linear regression. Furthermore, we establish a point of convergence on the transformation in the progression of users' (commenters') attitudes in ancient history over time, as well as the implications of the posts (queries based on different keywords) Linear regression can minimize squared errors and can be used in input-output training data sets.

CHAPTER III

System Model

3.1 Overview

In the above studies, we are ready to recognize tweets sentiments and studied in case or now not have difficulty to inventory markets. Large numbers of researches which are achievable to expect different market shares, in concluding expenses with corresponding to absorption, however, those extensively passes by European markets. On account of this, we end up with the conclusions by using discussed the DJIA values. This index includes top agencies of the world market with high-quality capitalization. Even as we need to confirm which forecast is tranquility affordable in a decreasing feed of tweets, although this is wising to comfy a comparatively massive core if anyone wants to practice applying different models. Obviously, many other international customers may additionally upload their opinions in English. By international stakeholders, taking them into attention in the sentiment analysis seems valid because the DJIA value is an influent.

Any other essential discussion and that have the benefit to be stated in lots of works and the opportunity of sentiment lexicons online. For positive, the most certain assets are accommodated in English. Then, collecting tweets and inventory information for the diverse timelines of different conditions. We are capable of counting every trend over the timeline and we keep away from any possible effect of bias giving thanks to seasonal volatility. An extended length additionally lets for validating the consequences with a higher degree of self-assurance. The execution approving which follows the general framework arranged to confirm inside the following discern. We take a look at a day-by-day statistic of tweets in parallel to a day-by-day inventory trade statistic.

After identifying the sentiment for everyday supported tweets content, we have been capable of trying to match a version to go seeking out the correlation. We have tried to put emphasis on assumption checking as most studying forget this step and without difficulty taking delivery of hypotheses. But we are going to perform regressions before checking the assumptions due to the fact we are going to use them in some manner to validate fine effects with a selected model shown in figure 3.1 instead of forsaking it at the same time as it can still provide heuristically desirable predictions.

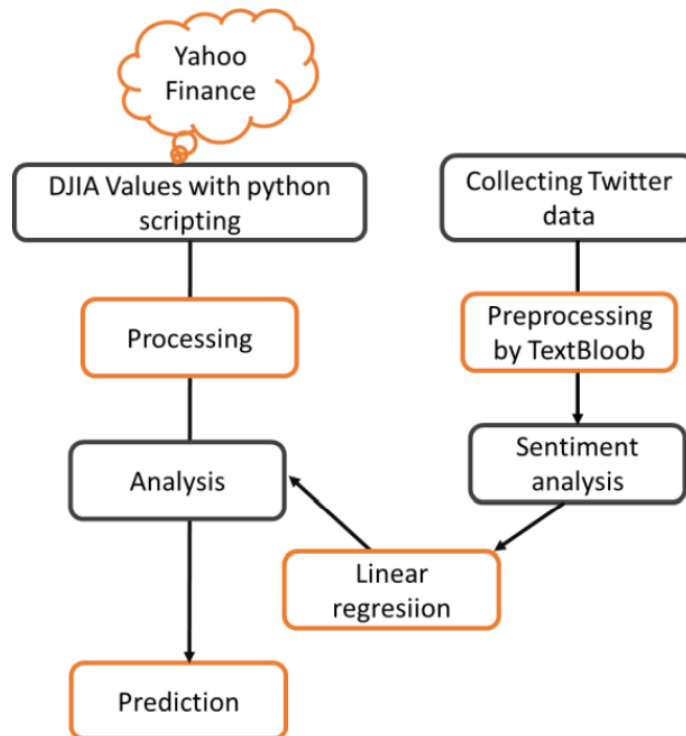


Figure 3.1: System Model Block Diagram

3.2 Methodology

While there was a lot of studies which was happened in classifying a bit of textual content as both positive or negative, there was little paintings on multi-magnificence classification. Sentiment analysis is a vital particle of our answer because this module of the output will be passed down for gaining knowledge of our predictive process. Natural language processing (NLP) is taken into consideration to be an effective device for expressing and decoding human languages like speech and textual content.

Natural language processing is beneficial to textual content evaluation and textual content mining. In our experiment, we use Python and TextBlob library to perform a variety of sports related to sentiment analysis. As Python programming library has easy API, TextBlob is effortlessly able to acting primary NLP sports and tasks.

3.3 System Design

In this project, the system we want to evaluate the sentiment analysis tool which retrieves, processes and tests twitter data. The main purpose of our project to build a co-relation between stock prices and the overall people sentiment. At the very first step we have collected stock values from Yahoo Finance for a defined time range for our project. Next step, we have tried to collect the twitter trending data using twitter API at the same time range. Then we have pre- processed data we got from twitter and DJIA values. With textblob and python data-frame for featuring twitter data for analyzing sentiment. At the final step there is involved a linear regression algorithm for visualizing of data along with required tests.the full process is shown as a system architecture in figure 3.2.

3.4 Data Set and Preprocessing

Because the ways of data collection can get in large portions, it has become difficult to identify from which source we extract data or exactly which messages are selected for doing sentiment analysis. Raw inventory price statistics are pre-processed before inputting into machine getting to know various models. Pre-processing consists of reworking the uncooked records within a layout which process can take from and function on, most possibly characteristic matrix. It allows extracting some features, monetary-area-particular especially, manually to enhance outcomes, allowing the model to research greater abstractions. The data we have collected for this research is mainly from 2020 till now.

In our study, we've collected random twitter data by using Twitter API, which we applied for. For the need of our experiments, we've collected random Twitter posts of that point as numeric data. These numeric data associated with finances are collected from Yahoo! Finance website. The latter is that the property of Yahoo!

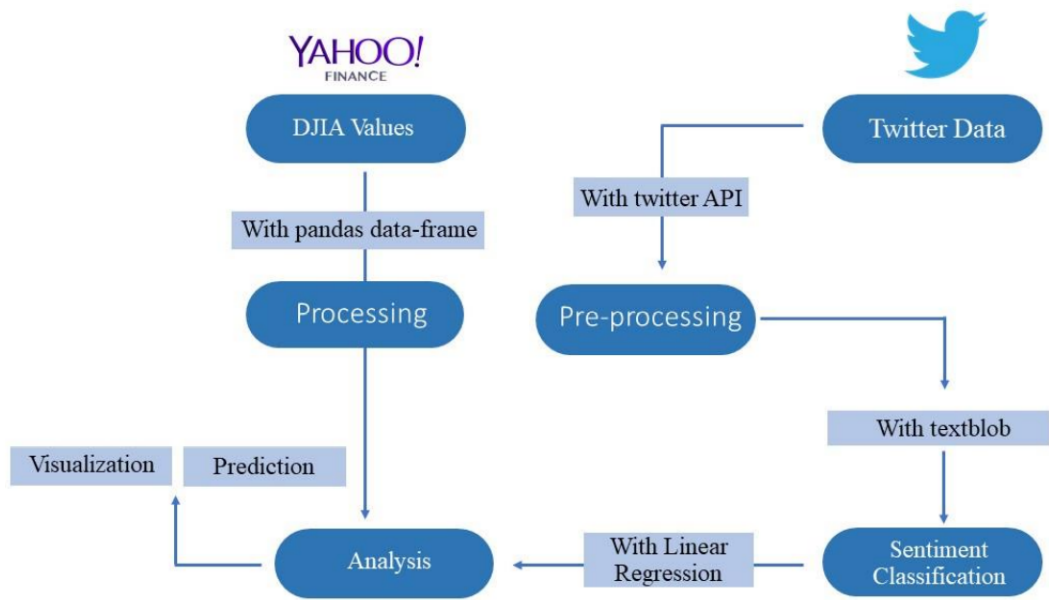


Figure 3.2: System Architecture

was founded to supply economic and financial news and commentary covering share prices, financial press releases, financial reports, etc. We collect each data while aggregating numeric data supported attributes of share prices with opening, high, low and shutting rates (OHLC). This data set is critical because it is employed to support predictive analyses on the future movement of share prices. The dataset processing is illustrated in figure 3.3.

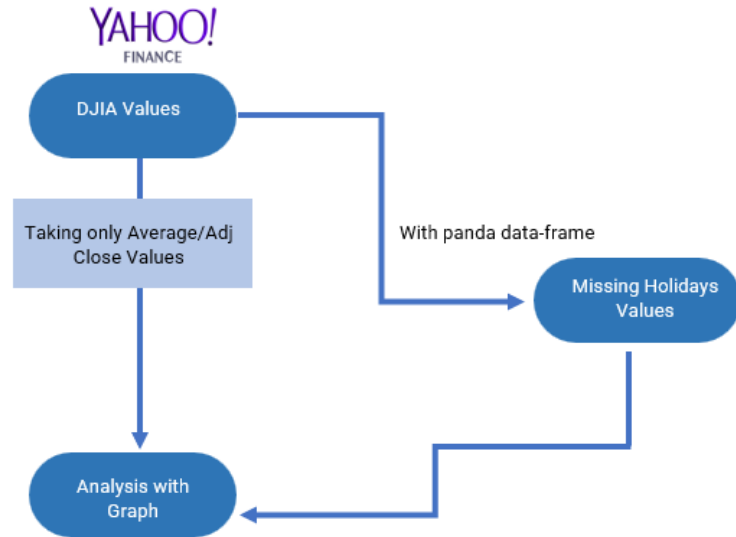
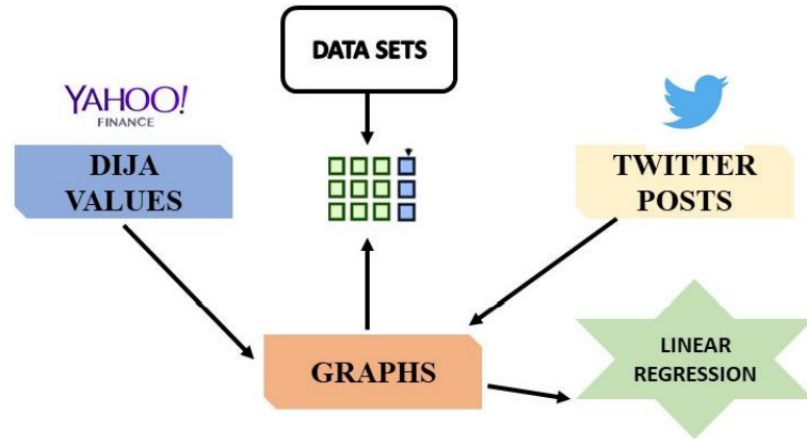


Figure 3.3: Dataset Pre-processing

3.4.1 DJIA Value

The main dataset for this study is the inventory stock data. For example, that index-wide variety industrial common (DJIA) is taken into consideration. To begin, create two variables: *start_date* and *end_date*. If each date found in yfinance, then the djia-values will be downloaded as a csv file, and finally the procedure has come to an end. Here, given below in the algorithm 1 which is used to extract DJIA data.

The primary information for the stock price records became to be had on the

Algorithm 1 DJIA Data Extract

```
1:  $each\_date \leftarrow (start\_date, end\_date)$ 
2:  $stock \leftarrow 'dji'$ 
3: for  $each\_date \in stock$  do
4:   if  $start\_date \leq end\_date$  then
5:      $djia\_values \leftarrow yfinance.download(stock, each\_date)$ 
6:   end if
7: end for
8:  $djia\_data \leftarrow pandas.dataframe(djia\_values)$ 
9: Return  $djia\_data$ 
```

Yahoo finance website. The know-how became accrued by writing a python script to carry out internet scraping.

	Date	Open	High	Low	Close	Adj Close	Volume
0	2019-01-02	37.470001	38.740002	37.410000	38.590000	32.369118	5537100
1	2019-01-03	38.959999	39.060001	38.480000	38.810001	32.553650	7137300
2	2019-01-04	39.570000	40.080002	39.419998	40.029999	33.576977	9336700
3	2019-01-07	39.700001	40.310001	39.580002	40.160000	33.686024	6114300
4	2019-01-08	40.180000	40.259998	39.830002	40.029999	33.576977	5654900
...
680	2021-09-14	25.400000	25.400000	24.580000	24.650000	24.650000	9315500
681	2021-09-15	25.379999	25.860001	25.360001	25.770000	25.770000	15378600
682	2021-09-16	25.730000	25.750000	25.230000	25.340000	25.340000	10188100
683	2021-09-17	25.420000	25.469999	24.959999	25.240000	25.240000	13286500
684	2021-09-20	24.940001	25.080000	24.530001	24.830000	24.830000	12743300

Figure 3.4: Collected DJIA Values

Through this net scraping shown in figure 3.4, the statistics are accumulated and stored as a comma-separated value (CSV) record. It's to be cited right here that most effective the inter-day buying and selling values are acquired. This refers back to the trading performed across diverse days and intra-day refers to the trade carried out in the day. This can be because the intraday trading fees don't appear to be readily available similar to the inter-day expenses and it also will increase the computational need and complexity. An additional piece of key statistics that may not be acquired is the order e-book. It can help provide a prediction of the rate using the weighted

average of the orders.

3.4.2 Twitter Data

Tweets are on hand via a truthful seek of requisite phrases via a utility programming interface (API). Currently, over 500 million messages are published on Twitter on an everyday foundation as we can see in figure 3.5. This takes a look at turned into carried out over a length of eight months length that is already instructed in the introduction. During this era, we're amassing English tweets wherein every tweet file contains various identifiers, languages, texts and different submitted dates/times to reach as we have directed our goal on trending tweets wherein tweets are different quite trades and mentioned generation stocks which have excessive tweet versions. [17]

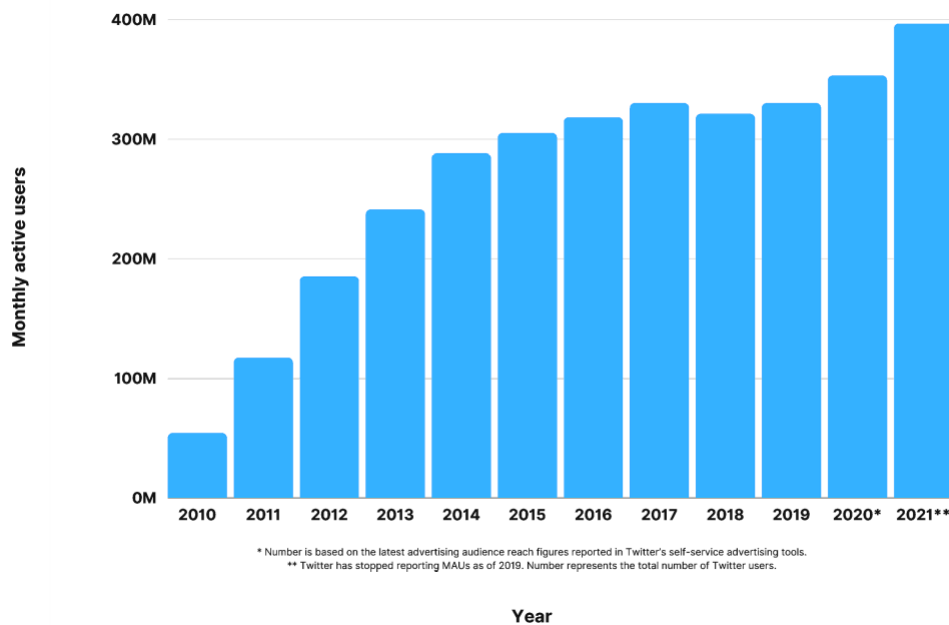


Figure 3.5: Twitter Users Growth

While the Twitter statistics became to be had for all day mendacity withinside the given timeline, the DJIA values are required the usage of Yahoo! Finance website became missing data for weekends and different desertions while that marketplace is shut. To finish these statistics, we resembled those lacking ethics the usage of pandas data-frame work.[24]

In the following steps how to obtain twitter API keys:

- Login to twitter developer section.
- Go to “Create an App”.
- Fill the details of the application.
- Click on Create your Twitter Application
- Details of your new app will be shown along with consumer key and consumer secret.
- For access token, click” Create my access token”. The page will refresh and generate access token.

After getting twitter API, we try to collect and extract twitter data. In given below algorithm which is used to help collecting and extracting twitter data.

To extract data from Twitter, we’ll need the API keys, which will allow us to start collecting data. We gained access to many keys after gaining access to the Twitter API. We use the following keys for data extraction: customer key, customer secret key, access key, access secret key, and access token key. We fixed these keys after initializing them as variables because they differed from user to user. We can get the author of any tweet using this API, which we use below with trending tweets available at different periods. Then, to get the available trending tweets on a specific day, we create a condition and put them into a try loop. If it doesn’t work, we’ll double-check the keys we received as well as our network connection. Then, in our trending data, we search for trending terms and select the most popular ones to create raw data, which we use to initialize for api search, q=word, and get the results of popular tweets. Finally, we save the tweet data as raw data for our collection using Pandas Frame. The algorithm 2 below, which is used to collect and extract Twitter data, is as follows:

Algorithm 2 Twitter Data Extract and Collect

Require: API Keys

```
1: tweet_data  $\leftarrow$  variableto collect tweet data
2: consumer_key  $\leftarrow$  tweetapi customer key
3: consumer_secret  $\leftarrow$  tweetapi customer secret key
4: access_key  $\leftarrow$  tweetapi customer access key
5: access_secret  $\leftarrow$  tweetapi customer secret access key
6: try
7:   auth  $\leftarrow$  tweepy.OAuthHandler(consumer_key, consumer_secret)
8:   auth.set_access_token(access_key, access_secret)
9:   api  $\leftarrow$  tweepy.API(auth)
10:  print('Authorized!')
11: except
12:  print('Error!')
13:
14: raw_data  $\leftarrow$  []
15: trending_data  $\leftarrow$  api.trends_place(id = 1)
16: for trending_word  $\in$  trending_data do
17:   for word  $\in$  trending_word['trends'] do
18:    raw_data  $\leftarrow$  tweepy.Cursor(api.search, q = word, type = 'popular')
19:   end for
20: end for
21: tweet_data  $\leftarrow$  pandas.dataframe(raw_data)
22: Return tweet_data
```

3.5 Sentiment Analysis

To categorize given contextual content in order to map it to elegance, sentiment analysis is used. Sentiment Analysis may be binary, i.e., high-quality or negative or multi-elegance with three or extra instructions involved. The sort of sentiment evaluation relies upon on dataset and reasoning method adopted. Researchers and stakeholders frequently diverge on their interpretation of courting that exists among sentiment evaluation and detection of emotion. While their factors of view may also fluctuate primarily based totally on their respective perspectives, researchers are united in adopting similar techniques. It is a manner to gauge or decide the emotion of the writer. The emotion can either be nice, bad or neutral. TextBlob's sentiment

characteristic evaluates to 2 properties polarity and subjectivity. Subjectivity belongings values also are flow that falls inside the variety of $[0, 1]$. How we have worked on our project with text blob are shown in the following figure 3.6:

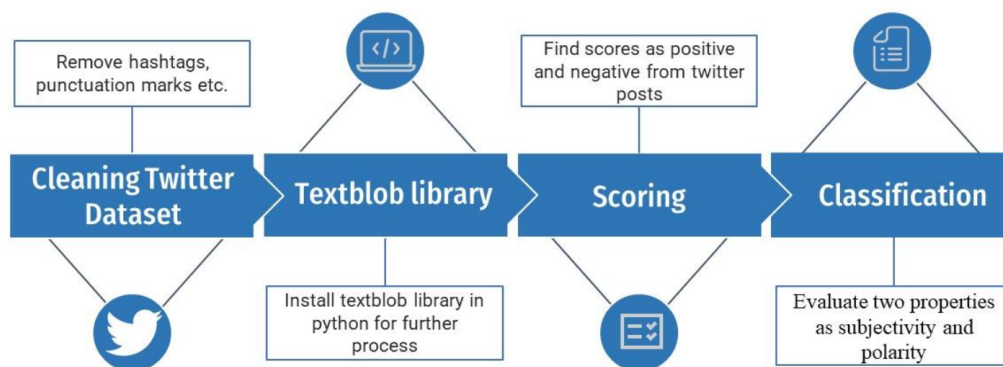


Figure 3.6: Textblob Working Process

In this study, the opinion of various peoples of various countries has been discussed. The most focus of this paper is on Twitter, Twitter API and have implemented the python artificial language and to implement the sentimental analysis as positive, negative and neutral. In order to make the Twitter API more relevant, we'll first import Tweepy. Then use tweet data, all tweets, tweet value, SE (set of emoji), SSC (set of special characters), SP (set of punctuations), and clean tweet as variables. Following that, we'll download tweets as datasets and save them to a CSV file. Additionally, the all-tweets variable is used to get tweets from the dataset, and the SE, SSC, and SP variables are used to remove and eliminate harmful emoji, #tags, special characters, and punctuation from tweets. Create a new CSV file for all clean tweets. The given algorithm helps us to organize and sanitize Twitter data.

We'll tokenize every phrase withinside the dataset and keep it into the dataset. For each phrase, will evaluate it with positive, terrible and impartial sentiments phrase withinside the dictionary. Then increment the positive, terrible and impartial count. Finally, based totally on the positive, terrible and impartial count, we are ready to get the top result percent approximately sentiment to see the polarity. This study suggests the sentimental evaluation set of rules at an excessive level. As may be visible withinside the algorithm 3, researchers have distinct approaches to attach the Twit-

Algorithm 3 Tweet Filtering Algorithm

Require: *tweet_data*

```
1: temoji  $\leftarrow$  setofEmojis
2: schar  $\leftarrow$  setofspecialcharacters
3: punc  $\leftarrow$  setofpunctuations
4:
5: raw_tweets  $\leftarrow$  tweet_data
6: for tweet  $\in$  raw_tweet do
7:   if emoji = schar = punc  $\neq$  [NULL] then
8:     clean_tweet  $\leftarrow$  remove  $\in$  emoji, schar, punc
9:   end if
10: end for
11: clean_data  $\leftarrow$  pandas.dataframe(clean_tweet)
12: Return clean_data
```

ter API, four fetch the tweets, tweet cleansing or do away with preventing phrases and punctuation marks, classify tweets because of this that get the polarity of the tweet, and finally go back the results. In this paper, python is used to put in force the sentimental evaluation. Some applications have been applied along with tweepy and textblob.[6] [7] The required libraries are mounted with the aid of using following commands:

- pip set up tweepy
- pip set up textblob

The main steps of this sentiment analysis are drawn in following figure 3.7:

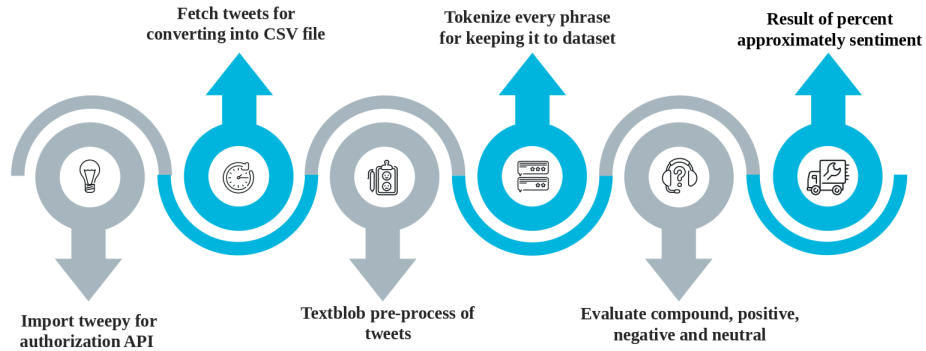


Figure 3.7: Sentiment Analysis Process

3.6 Hypothesis and Final Co-relation

For getting better result and prediction, we at first find out the p-value of our twitter data. We applied this co-efficient so that we can know whether the prediction of our data will be possible to get or not. A probability associated with a crucial value is known as a p-value. The crucial value is determined by the possibility of a Type mistake being allowed. It calculates the probability of achieving results that are as least as good as if the claim (H_0) were true.

A correlation matrix is a table that displays the coefficients of correlation between variables. The correlation between two variables is shown in each cell of the table. A correlation matrix can be used to summarize data, as an input to a more sophisticated study, or as a diagnostic tool for advanced analyses. The instances in a predicted class are represented by the rows of the matrix, whereas the instances in an actual class are represented by the columns (or vice versa). The name is derived from the fact that it is simple to observe if the system is mixing up two classes (i.e. commonly mis-labeling one as another). It's a unique type of contingency table with two dimensions ("polarity" and "subjectivity") with identical sets of "classes" in each dimension. As we know if p-value is stayed at between 0 to 0.1 it is possible to make prediction on that dataset. After applying p-value we got the average result 0.06, 0.07 and 0.3. We know the range of p value is -1 to 1. Since our result is 0.06, 0.07, 0.3 so, our dataset is ready for working. The p-value correlation coefficient of our dataset is given below in figure 3.8:

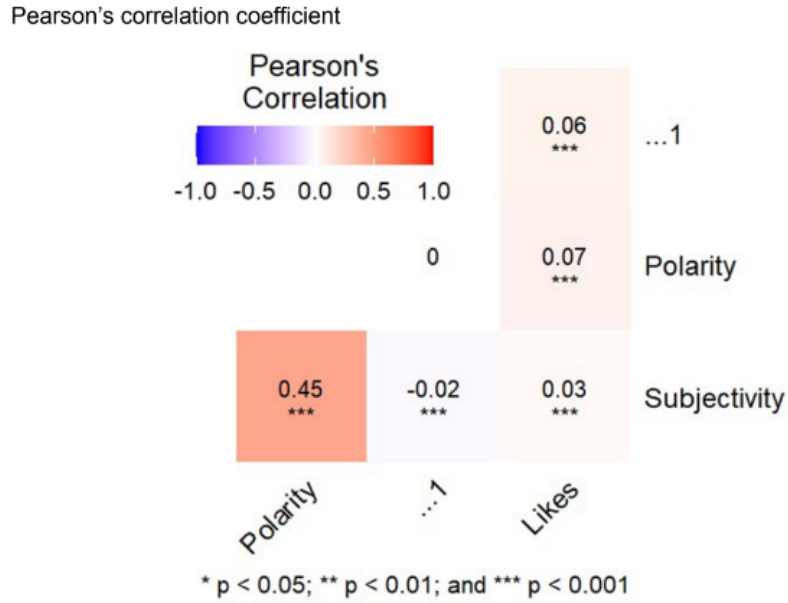


Figure 3.8: Pearson's correlation coefficient

This study will try to show the following hypothesis in order to analyze the market.

H1: The total averaged sentiment of all stocks within a sector will be used to determine the sentiment of the sector.

By using people sentiment, we will be able to identify their positive, negative thinking and whether their thinking is affecting a particular sector or not. The result of this hypothesis is clearly shown and explained in our result section with appropriate values and graph.

H2: On any given day, the sentiment of a sector or stock will provide a forecast for that stock's movement the next day.

We will know if the stock price will be up or down the next day or next moment of any day. By using this hypothesis, it will be easier for making any decisions to invest at their will. The overall prediction and accuracy level have been compared with our predicted values. To understand better it has been explained in detail at the result section with scatter plot by applying linear regression on our data.

H3: The stocks with the highest number of Twitter mentions are also the stocks that lend the most weight to a sector's mood.

From this thinking, we can know which type of sentiments are actually making effect

on stock prices. The sentiments of tweet will be clearly shown with related graph and values to understand. We will work on more data so that we can understand people's mood who are directly investing on share market. To know their mood, we need more data and also need to do real time survey on those people which will be time consuming so far for which we will work in future on that.

3.7 Linear Regression (Proposed Algorithm)

The linear regression set of rules tries to research a characteristic that maps enter vectors to scores. It represents with the aid of using a linear mixture of the enter capabilities. We will try to use an ahead step-clever technique to set the mass in order to carry out the characteristic collection. At the very early stage, we have to fix the implied evaluation rating. After this, till the favored variety of entering capabilities were delivered, enter capabilities have been delivered differentially. At every stride, that characteristic which will minimize corrupted mistakes maximum when delivered became collected. If the time of delivering, its mass fixes as if corrupted mistakes could be reduced, then given that capabilities and mass already delivered. That variety to enter capabilities consisting of became decided with the aid of using validation [17]. The working process of this algorithm is drawn at step by step in the following figure 3.9:

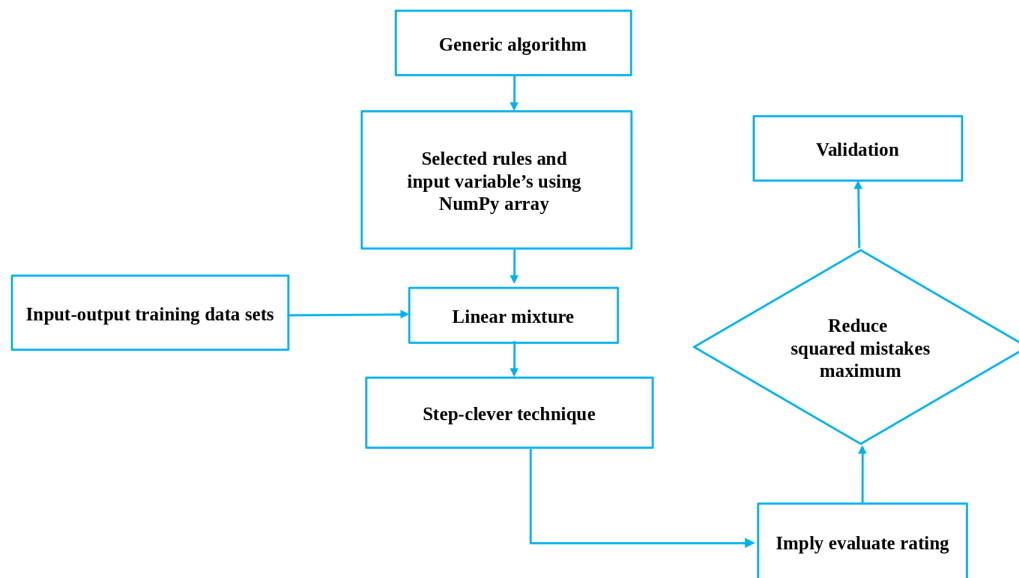


Figure 3.9: Linear Regression Workflow

CHAPTER IV

Result Analysis

After building all of our algorithms and drawing figures, we showed all the related results here with a prediction result in this segment.

We use pandas and yfinance to extract the DJIA value from Yahoo Finance. The yfinance module is a Python module that is very simple to use. The module 'yfinance' has grown in popularity as a python-friendly library that can be used as a patch to pandas datareader or as standalone library. It has a wide range of applications, and many people use it to download stock and cryptocurrency prices. Let's pretend we want to work with Google Finance's Close prices, which have already been updated to account for stock splits. We want to make sure that our dataset includes all weekdays, which is important for quantitative trading techniques. As a result, we get a graph drawn in figure 4.1 which providing DJIA value stock prices average.

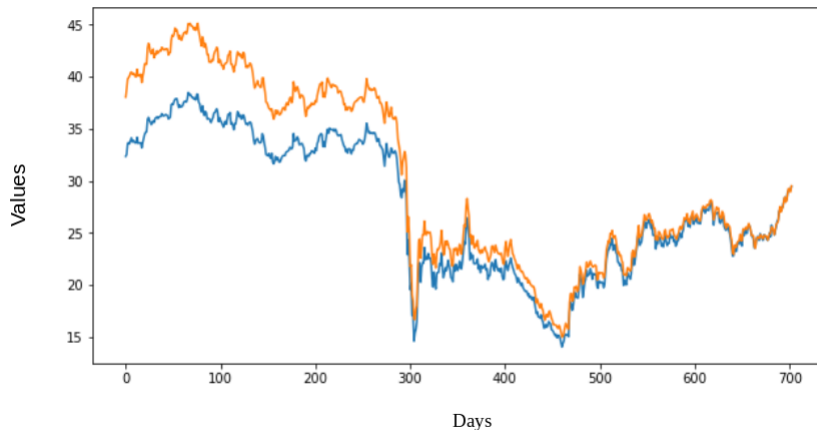


Figure 4.1: DJIA Values Stock Price Average

After getting DJIA values we collected our twitter data by using our twitter data extract and collect algorithm mentioned and explained in our system model. We extracted twitter data from 2019 till now and we added here some of our twitter data extracted in 2021 that is illustrated in figure 4.2.

































	2021_August_all_trends_data.csv	
	2021_August_hashtag_trend_data.csv	
	2021_August_twitter_trending_data.csv	
	2021_July_all_trends_data.csv	
	2021_July_hashtag_trend_data.csv	
	2021_July_twitter_trending_data.csv	
	2021_June_all_trends_data.csv	
	2021_June_hashtag_trend_data.csv	
	2021_June_twitter_trending_data.csv	
	2021_May_all_trends_data.csv	
	2021_May_hashtag_trend_data.csv	
	2021_May_twitter_trending_data.csv	
	2021_October_all_trends_data.csv	
	2021_October_hashtag_trend_data.csv	
	2021_October_twitter_trending_data.csv	
	2021_September_all_trends_data.csv	

Figure 4.2: Some Extracted Tweet

Then, we cleaned and filtered our extracted twitter data and stored in our drive for keeping it updated. We used twitter filtering algorithm and removed all types of punctuations, hashtags, emojis, tags which can be harmful to store raw data. These cleaned data will help to gain our fifth objective as every tweet is not important specially those tweets where special characters are added. We can see some of our sample cleaned tweets at the following figure 4.3 as shown below:

	Tweet
:38	Willz is becoming quite the baker https://t.co/6cMZuiGUBM
:24	This was so fun to watch for real thanks billboard https://t.co/aSSo8znrBA
:10	Do you mean promising NOT TO https://t.co/rbNusRK5bU
:10	Deal https://t.co/OMT2eF1zR
:44	I m fortunate because I ve never really depended on my looks I ve decided that my talent and my individuality is far more important than my face So get on board cause I
:35	I want my children to know what I look like when I m angry
:34	Continued note to self Every once in a while you consider altering your face and then you watch a show where you want to see what the person is feeling and their face i
:33	Letter to self Dear Me you re getting older I see lines Especially when you smile Your nose is getting bigger You look and feel weird as you get used to this new reality Bu
:03	Looks like we have a future Supercross in the making SupercrossLIVE https://t.co/oEIOJfRHAB
:13	Someone once told me that I too look like the fat version of pink Could be worse I guess https://t.co/YfZ5whMbVn
:45	At Anaheim tonight with my family and wishing Chad Reed the absolute best You are true perseverance and heart CRtwtotwo we love you
:49	My dreamy husband hartluck has been working hard to organize this super cool auction TanksForTroops via GoodRide A bunch of rad artists customized Indian motorcyc
:45	I am totally devastated watching what is happening in Australia right now with the horrific bushfires I am pledging a donation of 500 000 directly to the local fire services th
:49	I love Carol Burnett That is all
:40	Imagine all the people Living life in peace During this holiday season consider spreading peace love throughout the world by supporting the amazing life saving efforts c
:02	This made me cry real tears https://t.co/UjcyUyzVKA
14:3	All thanks to the fans friends and my awesome team https://t.co/7KWLw2U09H
:32	That s my girl https://t.co/wbzdnbRvQo
21:4	Wow https://t.co/Q6KXmkZh2q
17:4	realDonaldTrump What kind of grown man decides to write this tweet and then send it What kind of President tells a child to chill when she s trying to save the planet And i
15:0	President Trump on Thursday again publicly mocked teen climate crisis activist Greta Thunberg tweeting that the 16 year old has anger management issues Thunberg w
:11	Remember to be kind to the people in customer service the cashiers the people in retail your mailperson your delivery people They are working their asses off to make si
:47	My papa is a superhero like Spider Man https://t.co/Vqep6TxKXe
:08	No child should go hungry in America Join me in supporting NoKidHungry this GivingTuesday to help feed hungry youth Every 1 donated can provide up to 10 meals an
:39	This year I saw firsthand the work that unicefusa does for children in the hardest to reach places UNICEFWontStop until every child has what they need to thrive Join m
:47	I think my heart just stopped https://t.co/WxyXXVkdZcJ
:32	Yes She is a total badass and really great at what she does and I m very very happy that the world is recognizing her talents https://t.co/97Vf0jUNs1
:21	Nailed it https://t.co/vvO96dEDar

Figure 4.3: Sample Cleaned Tweets

With another machine learning and neural network approach like multiple regression, SVM, and others, our result will be able to provide the closest accuracy. We attempted to develop it for better output because Linear Regression is an old conventional model. The third objective of our work to predict by applying a proper algorithm text blob for sentiment. For the sentiment analyzer, we use textblob. After cleaning and filtering (removing hashtags, emoji, punctuation marks, and other obtrusive elements), we determine positive and negative scores from a collection of Twitter postings. Then use sentiment classification to assess subjectivity and polarity, two attributes. The float polarity which is in the range [-1,1], with 1 denoting a positive statement and -1 denoting a negative statement. Subjectivity statements which usually refer to personal feelings, emotions, or judgments, whereas objective sentences refer to facts. These graphs shown in figure 4.4 are offered to help us achieve our third goal of finding the best outcome.

```

      Polarity Compound
0      0.000000    0.0772
1      0.000000    0.0000
2      0.200000    0.4404
3      0.700000    0.4404
4      0.500000    0.0000
...
111351  0.000000    0.0000
111352  0.000000    0.1063
111353 -0.076250    0.7003
111354  0.114286    0.4019
111355 -0.050000   -0.1280

[111356 rows x 11 columns]>

```

Figure 4.4: Subjectivity and Polarity

On condition that our two number hypothesis, we have worked here with cleaned tweets and we have applied sentiment analyzer for getting positive, negative and neutral mood for a user. By analyzing or comparing we were able to show the effect of positive and negative moods specially on figure 4.5. From these values, we can determine any types of moods such as neutral, compound besides positive and negative moods of any users easily to prove our number one hypothesis.

Compound	Positive	Negative	Neutral
0.0772	0.053	0.000	0.947
0.0000	0.000	0.000	1.000
0.4404	0.420	0.000	0.580
0.4404	0.172	0.000	0.828
0.0000	0.000	0.000	1.000
...
0.0000	0.000	0.000	1.000
0.1063	0.094	0.073	0.833
0.7003	0.279	0.000	0.721
0.4019	0.074	0.000	0.926
-0.1280	0.000	0.067	0.933

Figure 4.5: Sentiment Analyzer values

To prove our first hypothesis, we have also shown a figure 4.6 which represents graph of our tweets on which we did our sentiment analyzing and tried to show it with different color for better understanding.

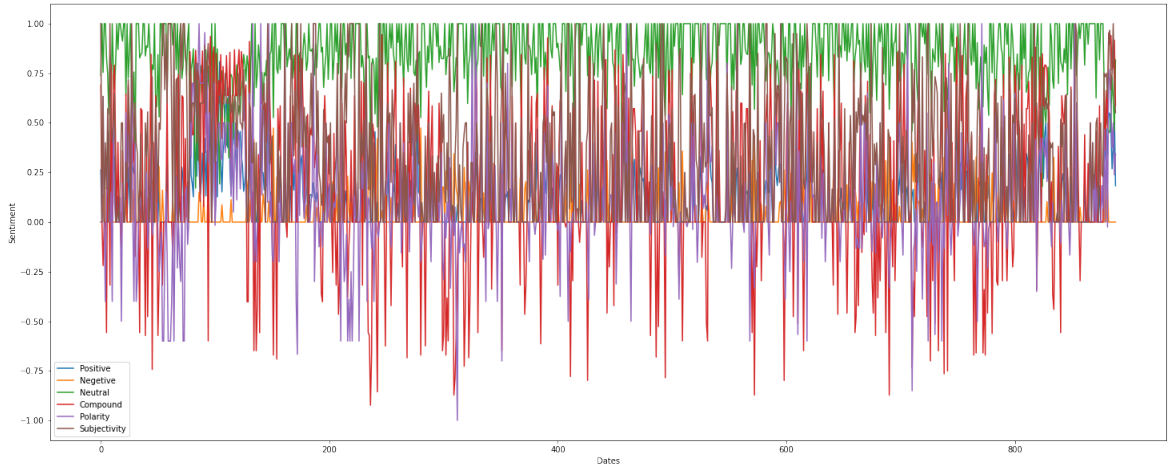


Figure 4.6: Sentiments of tweets

We have worked on our data that we acquired by applying the techniques indicated

above to gain an analysis on our project. After filtering and cleaning the raw data and deleting all forms of hashtags and punctuations, we ran our linear regression, which predicted whether the price of a share would rise or fall the next day with 73 percent accuracy. From the following figure 4.7 we can display our classification report of our model and this is highly used to show the precision, recall, f1-score and support on our trained data. This classification report is used to measure the performance evaluation in any machine learning model correctly. We also tried to show the mean absolute error so that we can be able to identify or measure how much our accuracy level is better. It fulfills our second objective that means it is reducing complexity more than before and also giving higher accuracy approximately 73% which is already mentioned. Also, this prediction helped us to prove our second hypothesis clearly.

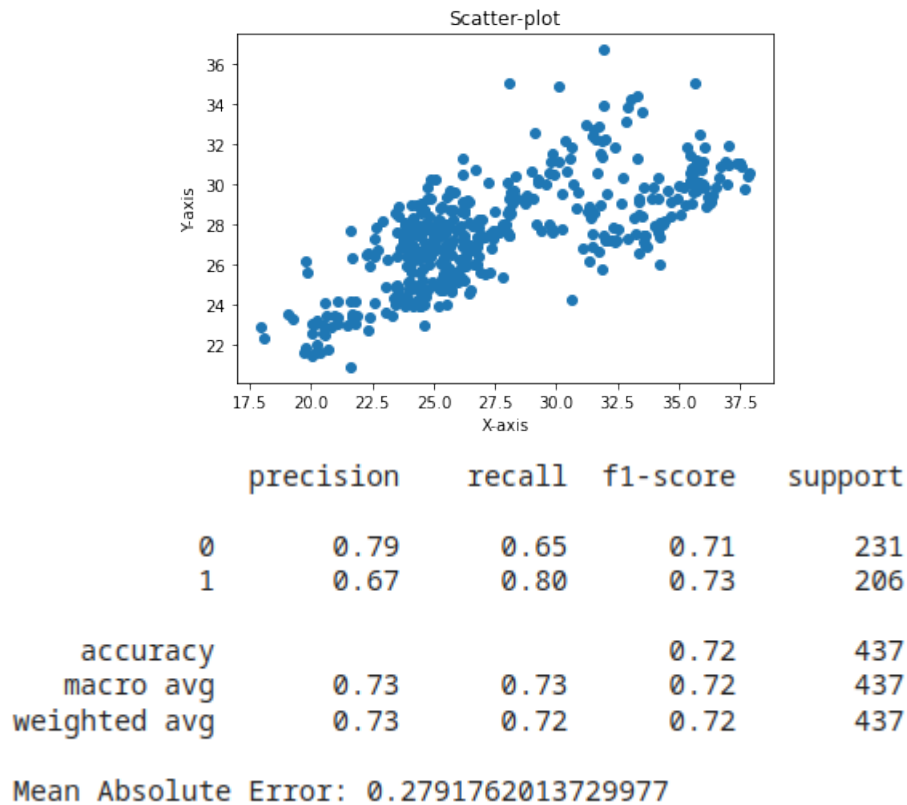


Figure 4.7: Predicted Percentage

The scatter plot as shown in the figure 4.7 represents the prediction values vs real values plot where we can show the difference. Also, from the below figure we can see clearly the difference between the actual and predicted values.

	Actual	Predicted	Difference
0	29.953773	31.647317	-1.693543
1	26.613913	25.772170	0.841743
2	23.872141	27.669325	-3.797185
3	29.201887	32.243247	-3.041360
4	21.703274	23.184932	-1.481658
...
650	20.981871	19.983720	0.998150
651	33.216988	28.133330	5.083658
652	25.458603	28.003673	-2.545070
653	24.892557	28.126110	-3.233553
654	26.378479	27.942551	-1.564072

655 rows × 3 columns

Figure 4.8: Actual vs Predicted

From the above figure 4.8 we can see that the difference between actual and predicted value is almost close. There is no huge difference between them as we can see. Also, we showed this difference in a graph with a green and red color where actual values are seemed to green and predicted values are shown as a red color. Thorough this figure 4.9 which illustrates a graph, any one can be able to see the difference between them.

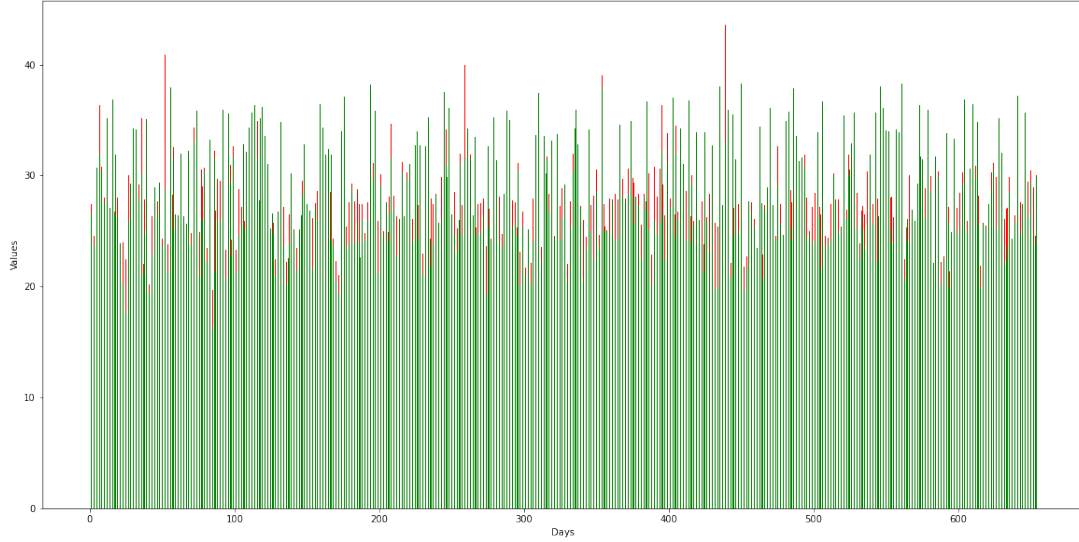


Figure 4.9: Actual values vs Predicted Values

Any user will be able to decide whether or not to acquire the stock, and this will be achievable if there are good sentiments; however, if the mood remains negative, the stock may need to be sold. With another machine learning and neural network approach like multiple regression, SVM, and others, our result will be able to provide the closest accuracy. We attempted to develop it for better output because Linear Regression is an old conventional model.

4.1 Comparison with Other Works

In this segment, we compare our result with other projects at the following table so that we can show where we have been able to improve. From the research gap section, we compared their accuracy, approach in detail and also discussed their drawbacks in which they can improve their projects and also added our gap which can be developed in future.

Titlle of Works	Prediction of stock values using sentiment analysis on twitter data [22]	Examining the effects of pan-demics on stock market trends through sentiment analysis[23]	Titlle of our work ”Stock Market Analysis Using Twitter Sentiment”
Approach	The model that processes live tweets and categorizes them as good, negative, or neutral, based on Twitter data	A variety of news stories with numerical values in order to better comprehend stock market trends based on previous patterns	Trending Tweets analyse with textblob and sentiment analyzer for sentiment analysis
Results	67% directional accuracy for DJIA	50% accuracy of positive and negative value	73% accuracy for DJIA of positive, negative, compound, neutral mood
Feedback/ Drawbacks	The results would be less accurate, hinting that the emotion of the people they analyzed isn’t totally accurate.	The use of aggregated share prices and the lack of daily news and articles gathered from different financial and stock market websites	The proposed dataset does not represent true public mood; it only takes into account persons who use Twitter and speak English., it is feasible to establish a higher correlation.

CHAPTER V

Conclusion and Future scope

Using data from Yahoo Finance, this research constructs a model and algorithm for stock price prediction. Traders, investors, and analysts benefit from efficient and accurate stock price prediction systems since they provide helpful information such as the stock market's future direction. For the best accuracy, we discovered that historical data should be used for Linear Regression. Data scientists' possibilities for solving fascinating problems with high accuracy are expanding thanks to deep learning technologies. We used a large scale collection of tweet data to uncover links between Twitter-based sentiment analysis of a particular company/index and its market performance in this article. Our findings reveal that negative and positive aspects of public mood have a strong cause-and-effect relationship with individual stock/index price changes. We also looked into how prior week sentiment features influence the opening and closing values of stock indexes for various tech businesses and important indexes such as the DJIA and NASDAQ-100. Our findings are in some ways similar, but there are some significant variances. To begin, our findings reveal a stronger association between negative and positive mood dimensions and DJIA values, as opposed to their findings, which showed a strong correlation with only the good mood dimension. Second, while we were unable to achieve a high percentage result of 87 percent, our 73 percent result using Linear regression provides better proof that the connection exists over the full data set.

Finally, numerous factors are left out of our analysis. To begin with, our dataset does not represent true public mood; it only takes into account persons who use Twitter and speak English. When the actual mood is analyzed, it is feasible to establish a higher correlation. The association could be explained by the fact that people's moods influence their financial decisions. However, there is no clear link between persons who invest in stocks and those who use Twitter more regularly, however there

is an indirect link: people's investment decisions may be influenced by the emotions of those around them, i.e. general public attitude. All of these will continue to be investigated in the future.

References

- [1] V. K. S. Reddy, “Stock market prediction using twitter sentiment analysis,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 08, no. 10, 2021.
- [2] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro, “How useful are your comments? analyzing and predicting youtube comments and comment ratings,” in *Proceedings of the 19th international conference on World wide web*, pp. 891–900, 2010.
- [3] A. Sureka, P. Kumaraguru, A. Goyal, and S. Chhabra, “Mining youtube to discover extremist videos, users and hidden communities,” in *Asia information retrieval symposium*, pp. 13–24, Springer, 2010.
- [4] F. Cheong and C. Cheong, “Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods.,” *PACIS*, vol. 11, pp. 46–46, 2011.
- [5] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross, “Identifying video spammers in online social networks,” in *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pp. 45–52, 2008.
- [6] N. Bhandari, “Stock market trend prediction using sentiment analysis,” 2017.
- [7] A. Mittal and A. Goel, “Stock prediction using twitter sentiment analysis,” *Stanford University, CS229 (2011 [http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf](http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf))*, vol. 15, p. 2352, 2012.

- [8] A. Krishna, J. Zambreno, and S. Krishnan, “Polarity trend analysis of public sentiment on youtube,” in *Proceedings of the 19th international conference on management of data*, pp. 125–128, 2013.
- [9] L. Nemes and A. Kiss, “Prediction of stock values changes using sentiment analysis of stock news headlines,” *Journal of Information and Telecommunication*, vol. 5, no. 3, pp. 375–394, 2021.
- [10] D. Shah, H. Isah, and F. Zulkernine, “Predicting the effects of news sentiments on the stock market,” in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 4705–4708, IEEE, 2018.
- [11] D. Peng, “Analysis of investor sentiment and stock market volatility trend based on big data strategy,” in *2019 International Conference on Robots & Intelligent System (ICRIS)*, pp. 269–272, IEEE, 2019.
- [12] Y. E. Cakra and B. D. Trisedya, “Stock price prediction using linear regression based on sentiment analysis,” in *2015 international conference on advanced computer science and information systems (ICACSIS)*, pp. 147–154, IEEE, 2015.
- [13] V. Gururaj, V. Shriya, and K. Ashwini, “Stock market prediction using linear regression and support vector machines,” *Int. J. Appl. Eng. Res*, vol. 14, no. 8, pp. 1931–1934, 2019.
- [14] D. Valle-Cruz, V. Fernandez-Cortez, A. López-Chau, and R. Sandoval-Almazán, “Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the h1n1 and the covid-19 periods,” *Cognitive computation*, pp. 1–16, 2021.
- [15] C. Gondaliya, A. Patel, and T. Shah, “Sentiment analysis and prediction of indian stock market amid covid-19 pandemic,” in *IOP Conference Series: Materials Science and Engineering*, vol. 1020, p. 012023, IOP Publishing, 2021.
- [16] X. Wang and X. Luo, “Sentimental space based analysis of user personalized sentiments,” in *2013 ninth international conference on semantics, knowledge and grids*, pp. 151–156, IEEE, 2013.
- [17] A. Bhardwaj, Y. Narayan, M. Dutta, *et al.*, “Sentiment analysis for indian stock market prediction using sensex and nifty,” *Procedia computer science*, vol. 70, pp. 85–91, 2015.

- [18] S. Papadamou, A. Fassas, D. Kenourgios, and D. Dimitriou, “Direct and indirect effects of covid-19 pandemic on implied stock market volatility: Evidence from panel data analysis,” 2020.
- [19] V. K. S. Reddy, “Stock market prediction using machine learning,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 10, pp. 1033–1035, 2018.
- [20] T. H. Nguyen and K. Shirai, “Topic modeling based sentiment analysis on social media for stock market prediction,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1354–1364, 2015.
- [21] P. K. Singh, A. Sachdeva, D. Mahajan, N. Pande, and A. Sharma, “An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites,” in *2014 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence)*, pp. 329–335, IEEE, 2014.
- [22] A. .V and D. V. Priya, “Prediction of stock values using sentiment analysis on twitter data,” *International Journal of Science and Research (IJSR)*, 2018.
- [23] S. Biswas, I. Sarkar, P. Das, R. Bose, and S. Roy, “Examining the effects of pandemics on stock market trends through sentiment analysis,” *Journal of Xidian University*, vol. 14, no. 6, pp. 1163–1176, 2020.
- [24] “Count NaN or missing values in Pandas DataFrame,” July 2020. <https://www.geeksforgeeks.org/count-nan-or-missing-values-in-pandas-dataframe/>.