

# **STOCK MARKET PRICE PREDICTION USING SENTIMENT ANALYSIS**

*Project report submitted in partial fulfilment of the requirement of the degree of*

**BACHELOR OF TECHNOLOGY**

**IN**

**ELECTRONICS AND COMMUNICATION ENGINEERING**

By

**Kritik Verma (171009)**

**Keshav Vohra (171016)**

**UNDER THE GUIDANCE OF**

**Dr. Naveen Jaglan**



**JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,  
WAKNAGHAT**

**December 2020**

# TABLE OF CONTENTS

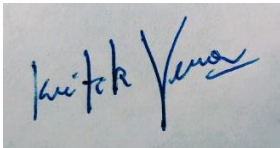
<b>CAPTION</b>	<b>PAGE NO.</b>
DECLARATION	i
ACKNOWLEDGEMENT	ii
LIST OF ACRONYMS AND ABBREVIATIONS	iii
LIST OF FIGURES	iv
ABSTRACT	vi
<b>CHAPTER-1: INTRODUCTION</b>	<b>1</b>
1.1 Stock Market	1
1.2 Sentiment Analysis	2
1.3 Problem Statement	3
1.4 Procedure	5
1.5 Classification Algorithms	5
1.5.1 Logistic Regression	5
1.5.2 Naive Bayes	5
1.5.3 K-Nearest Neighbours	7
1.5.4 Support Vector Machine	7
<b>CHAPTER-2: LITERATURE SURVEY</b>	<b>8</b>
2.1 Prediction Models for Indian Stock Market	8

2.1.1 Proposed Method	7
2.1.2 Conclusion	10
2.2 Stock Market Prediction Using Machine Learning	10
2.2.1 Proposed Method	11
2.2.2 Conclusion	12
<b>CHAPTER-3: SENTIMENT ANALYSIS OF DIFFERENT COMPANIES</b>	<b>14</b>
3.1 Reliance	14
3.1.1 Results	14
3.1.2 Conclusion	16
3.2 Tata Consultancy Services	16
3.2.1 Results	16
3.2.2 Conclusion	18
3.3 Wipro	19
3.3.1 Results	19
3.3.2 Conclusion	21
3.4 Infosys	21
3.4.1 Results	22
3.4.2 Conclusion	23
<b>CHAPTER-4: STOCK MARKET PRICE PREDICTION USING TEXTBLOB FOR SENTIMENT ANALYSIS</b>	<b>24</b>
4.1 TextBlob	24
4.2 Code	24

4.3 Results	29
<b>CHAPTER-5: STOCK MARKET PRICE PREDICTION USING NLTK FOR SENTIMENT ANALYSIS</b>	<b>30</b>
5.1 NLTK (Natural Language Tool Kit)	30
5.2 Code	30
5.3 Results	35
<b>CHAPTER – 6: CONCLUSION</b>	<b>37</b>
<b>REFERENCES</b>	<b>38</b>

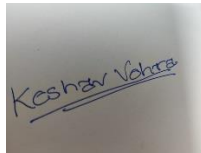
## DECLARATION

We hereby declare that the work reported in the B.Tech Project Report entitled “**Stock Market Price Prediction Using Sentiment Analysis**” submitted at **Jaypee University of Information Technology, Wazirpur, India** is an authentic record of our work carried out under the supervision of Dr Naveen Jaglan. We have not submitted this work elsewhere for any other degree or diploma.



Kritik Verma

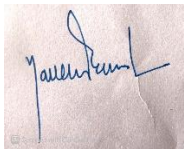
171009



Keshav Vohra

171016

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.



Dr Naveen Jaglan

Date:

Head of the Department/Project Coordinator

## **ACKNOWLEDGEMENT**

We would like to thank God for guiding us throughout our academic journey and to acknowledge our project supervisor, Dr Naveen Jaglan, for his undying support, priceless motivation and guidance throughout the project duration. Moreover, we extend our sincere gratitude to all the faculties and non-teaching staff of the Department of Electronics and Communication Engineering for their contribution towards the success of this work.

The role our friends played during the entire period cannot also go unmentioned. Thank you all for your moral support and encouragement. We deeply honoured and indebted to you all.

To our families, we appreciate the support you have given us throughout our academic journey. This quest has not been easy but you have always solemnly stood by our side.

Thank you.

## LIST OF ACRONYMS AND ABBREVIATIONS

CAGR	Compound Annual Growth Rate
API	Application Programming Interface
OS	Operating System
AI	Artificial Intelligence
K-NN	K NearestNeighbor
SVM	Support Vector Machine
ML	Machine Learning
CSV	Comma Separated Values
GUI	Graphical User Interface
IT	Information Technology
ADR	Association of Democratic Reforms
NASDAQ	National Association of Securities Dealers Automated Quotations
BSE	Bombay Stock Exchange
FICO	Fair Isaac Corporation
NLTK	Natural Language Tool Kit

## LIST OF FIGURES

Figure 1.1: Stock Exchange.....	1
Figure 1.2: Sentiment Analysis.....	2
Figure 1.3: Electoral College projections for 2020 US elections using sentiment analysis.....	3
Figure 1.4: Global Sentiment Analysis in Software Market.....	4
Figure 1.5: Sigmoid Function used in Logistic Regression.....	5
Figure 1.6: Naïve Bayes Theorem.....	5
Figure 1.7:K-NN Classification.....	6
Figure 1.8: Support Vector Machine Classification.....	6
Figure 2.1: Flow Chart for Daily Prediction Model.....	8
Figure 2.2: Learning Environment.....	10
Figure 3.1: Pie Chart for Sentiment Analysis on Reliance.....	12
Figure 3.2: Bar Graph for Sentiment Analysis on Reliance.....	13
Figure 3.3: Line Graph for Sentiment Analysis on Reliance.....	13
Figure 3.4: Pie Chart for Sentiment Analysis on TCS.....	15
Figure 3.5: Bar Graph for Sentiment Analysis on TCS.....	15
Figure 3.6: Line Graph for Sentiment Analysis on TCS.....	16
Figure 3.7: Pie Chart for Sentiment Analysis on Wipro.....	17
Figure 3.8: Bar Graph for Sentiment Analysis on Wipro.....	18
Figure 3.9: Line Graph for Sentiment Analysis on Wipro.....	18
Figure 3.10: Pie Chart for Sentiment Analysis on Infosys.....	20
Figure 3.11: Bar Graph for Sentiment Analysis on Infosys.....	20



Figure 3.12: Line Graph for Sentiment Analysis on Infosys.....	21
Figure 4.1: Importing Necessary Libraries.....	24
Figure 4.2: Initializing and Authenticating Keys and Tokens.....	25
Figure 4.3: Getting the Tweets.....	25
Figure 4.4: Cleaning the Data.....	26
Figure 4.5: Sentiment Analysis.....	27
Figure 4.6: Clubbing the Data.....	28
Figure 4.7: Getting the Stock Data.....	28
Figure 4.8: Training and Testing.....	29
Figure 5.1: Importing Necessary Libraries.....	30
Figure 5.2: Reading the Scraped Data.....	31
Figure 5.3: Cleaning the Data.....	31
Figure 5.4: Sentiment Analysis Using NLTK.....	32
Figure 5.5: Clubbing the Data.....	33
Figure 5.6: Retrieving the Historical Stock Data.....	33
Figure 5.7: Training and Testing.....	34
Figure 5.8: Linear Regression.....	34
Figure 5.9: Ridge Regression.....	35
Figure 5.10: Lasso Regression.....	35

## **ABSTRACT**

The intent of this project is to create a stock market prediction model based on the current trends in the market place, historical data and the general sentiment of the public based on different media outlets.

Financial market value data is created in gigantic volume and it just changes in a second. Stock market is an intricate and testing framework where individuals will either pick up cash or lose as long as they can remember investment funds. In this work, an endeavor is made for forecast of stock market pattern. Two models are constructed one for every day expectation and the other one is for month to month forecast. In this we already have the data so we can create supervised model and also test it. Model with best accuracy can be chosen.

Sentiment investigation is relevant extracting of text which distinguishes also removes abstract data in source material, and helping a business to know the social sentiment of their image, item or administration while checking on the web forums. Sentiment examination (or assessment mining) utilizes normal language preparing and AI to decipher and group feelings in abstract information. Up to 70% of precision is noticed utilizing regulated AI calculations on every day expectation model. Month to month forecast model attempts to assess whether there is any comparability between any two months pattern. Assessment demonstrates that pattern of one month is least related with the pattern of one more month.

For this project the coding is done in Jupyter Notebook.

# CHAPTER – 1

## INTRODUCTION

### 1.1 Stock Market

A stock market, equity market or share market, where the promoter of the company raises money from general public basically it can also be raised by banks, mutual funds, QIB investors. For investing in stock market we just need a demat and trading account affiliated with a stock broker. In our country there are two main stock exchanges, NSE and BSE where almost 1600 companies are listed. Large volumes of shares are traded on every market day. More the volatility in the market, faster it changes.



**Fig 1.1:** Stock Exchange

In simple words, in stock exchange is an open market place where we can buy and sell the shares of the publically traded company it is regulated by SEBI.

Not a lot of people prefer to invest their money in stock market because of the high risk that is involved in it but those who dare can get good to excellent returns. In India alone, where the population is 1.3 billion, there are only 18 million investors in equity market.

Before jumping blindly into investing people tend to check various statistics, graphs, charts, etc. in

order to get a good read on where to invest and where to not invest. A general public sentiment plays a huge role on whether a stock will perform well or not. Keeping in mind that this is not the only factor that might affect the final result but it can be considered as a major factor. In this project we are just focussing on the sentiment analysis and how it will affect the final prediction of a stock.

## **1.2 Sentiment Analysis**

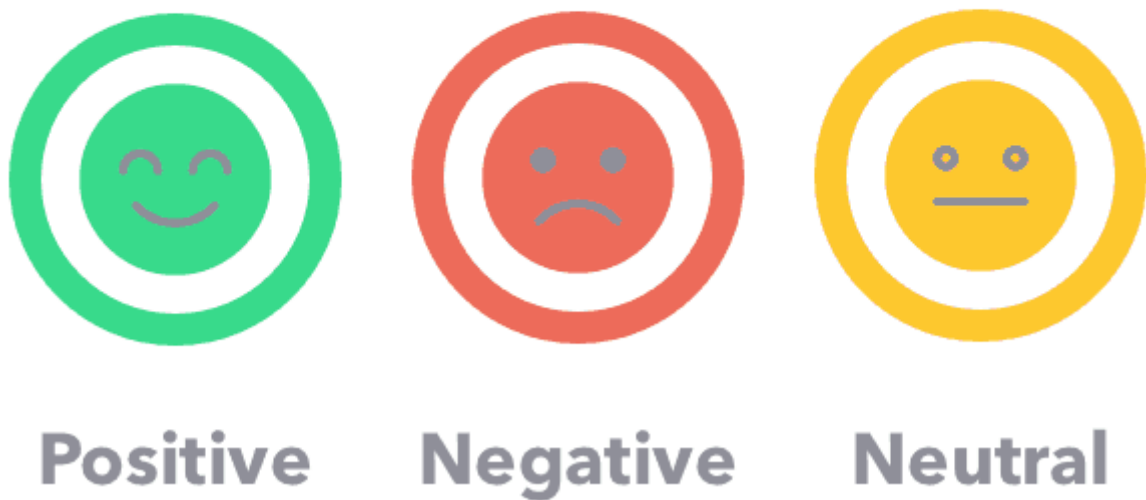
Basically, assumption investigation or assessment grouping fall into the general class of text arrangement undertakings where you are provided with an expression, or a rundown of expressions and your classifier should tell if the estimation behind that is positive, negative or impartial. Once in a while, the third characteristic isn't taken to keep it a parallel characterization issue. In ongoing errands, slants like "fairly certain" and "to some degree negative" are additionally being thought off.

As wistful investigation has improved over the most recent couple of many years so has its applicatons.Sentimental analysis[1] is presently being utilized from explicit item showcasing to hostile to s social conduct acknowledgment. Nowadays there are many medium from which general public can show their emotions like through Facebook, twitter, youtube, and the other small websites .In earlier days there were only news channels were the medium now every one show their point of view .

Sentimental analysis is not just seeing the emotions for stock market it can be checked for any topic like elections. Nowadays every political party have account on social media so it very important tool in elections. Since in this digital age voices can reach farther through this.

As more and more users post about products and services they use, or express their political and religious views on internet it becomes valuable information because from that analyst can predict opinion of general public and then there can be a target advertisement which turn to be extremely profitable [2].

# Sentiment Analysis



**Fig 1.2:** Sentiment Analysis

## 1.3 Problem Statement

To gather data about specific product from Internet(twitter,Instagram,News sites etc) . make general opinion about that.

This could be advantageous over the accompanying territories:

### a) **Business**

Nowadays best way to promote any business is social media or through digital marketing. Business usually already have our data through Google or social they our continuously monitoring our interest , so that's how they can target the right audience.

### b) **Politics**

In political field, social media is great tool through this parties could analyse public view area wise since mostly all the applications ask permission for our locations so it much better way than calling and asking voters about their vote.

### c) **Reviews and Ratings**

Online platforms use many types of classification algorithms through which they determine

the right product for you .On many online websites there is customer review section in which they analyse all the customer reviews , Not ecommerce sites like amazon uses it but also OTT platform like Netflix uses it they already know what thing you like to watch so they specifically show relatable stuff only.

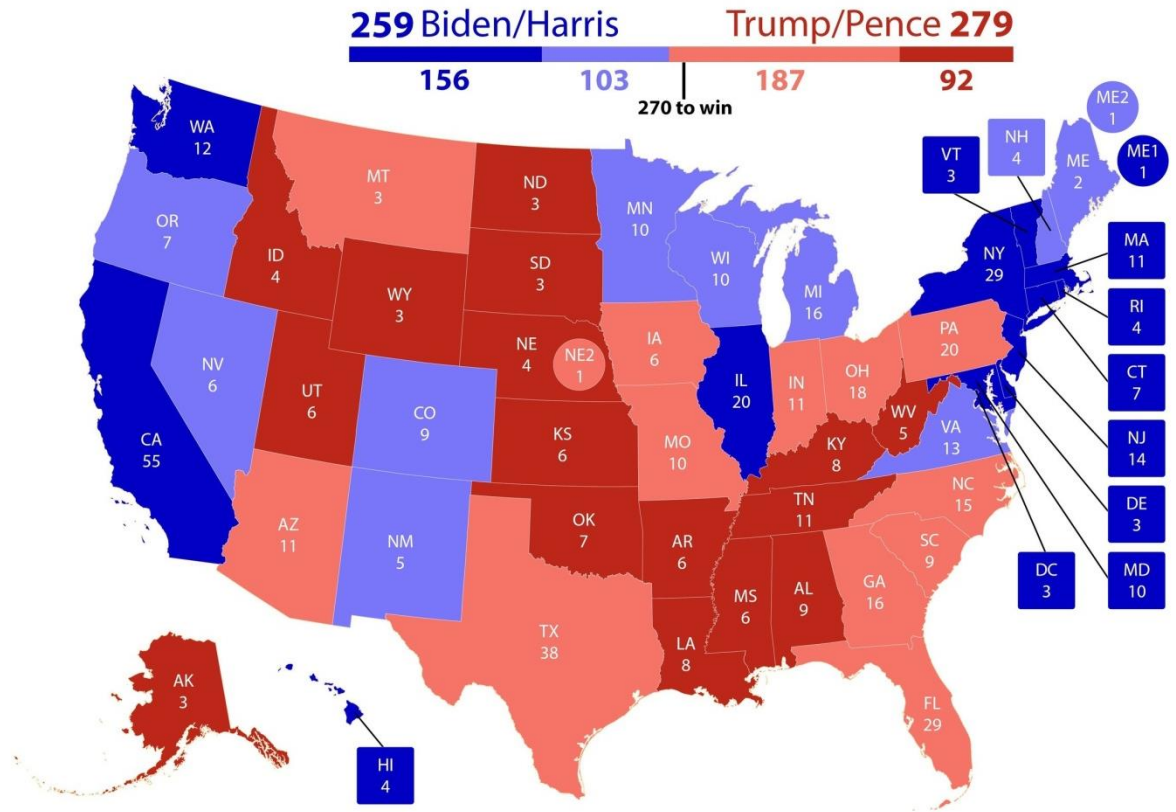


Fig 1.3: Electoral College projections for 2020 US elections using sentiment analysis[3]

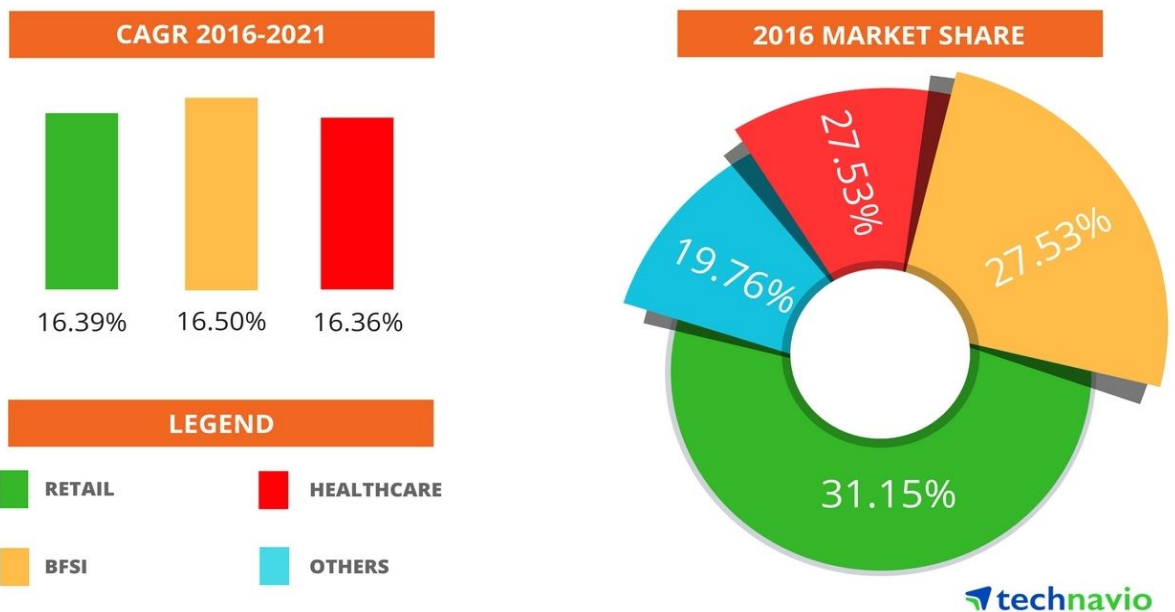


Fig 1.4: Global Sentiment Analysis in Software Market

## **1.4 Procedure**

Using “Tweepy” which is official library provided twitter collects tweets about specific subject. The removed information is investigated through python at runtime, by methods for numerical capacities for the normal just as standard deviation the variety of specific #Hashtag or "Search Term" is determined and the outcomes are put away in a "<searchword>.csv" document (that too at runtime, if the record doesn't exist beforehand, the O.S. progressively makes on at run time).

With the end goal of representation of information the matplotlib library is utilized which helps in indicating the 2-D plot for given number of days.

## **1.5 Classification Algorithms**

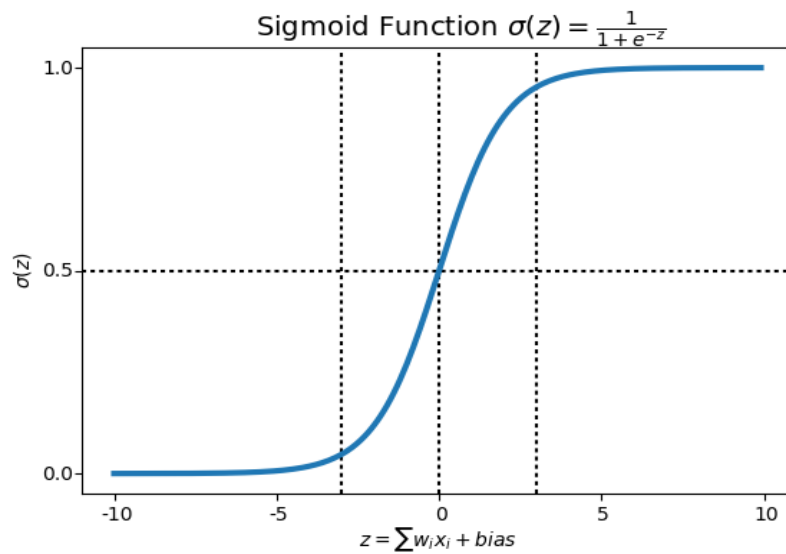
Listed below are a few techniques that can be used in order to perform the above procedures of Sentiment analysis and future value prediction.

### **1.5.1 Logistic Regression**

Logistic Regression is a classification algorithm which uses a special function known as sigmoid function it divides the problem statement in binary way (0 or 1) .

Now days it used in many places such as company recruitment many companies uses this to find the right candidate.

assumes data is free of missing values.

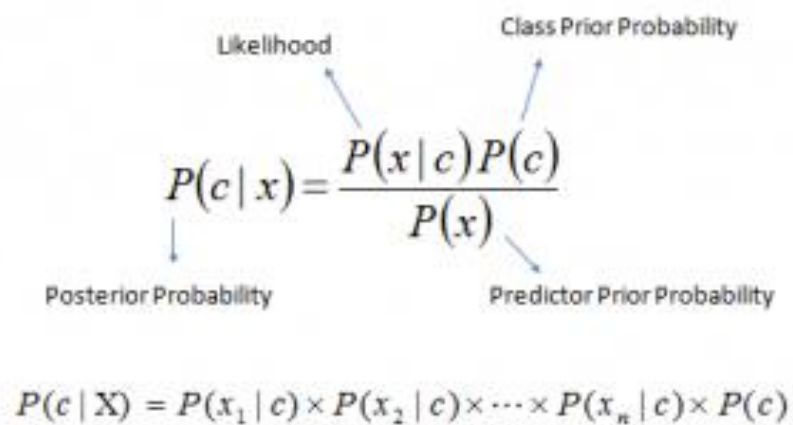


**Fig 1.5:** Sigmoid Function used in Logistic Regression

### 1.5.2 Naïve Bayes

Naive Bayes calculation is dependent on Bayes' hypothesis with the suspicion of freedom between each pair of highlights.

Naive Bayes classifiers work well as it is faster than other classifiers and it used in many areas such as G-Mail uses it for spam mail filtration and it also used for search engine optimization(SEO).



**Fig 1.6:** Naïve Bayes Theorem



### 1.5.3 K-Nearest Neighbors

K-Nearest Neighbors is a type of supervised machine learning algorithm which can be used as both classifier and for regression. In KNN we use a lot of labeled data such as, we recognize “book” we say book, if it is not a book “no book”. It uses for image recognition also. In the algorithm to the pointer we find the nearest partition by the distance it is also known as k-value. The partition to pointer minimum distance is chosen.

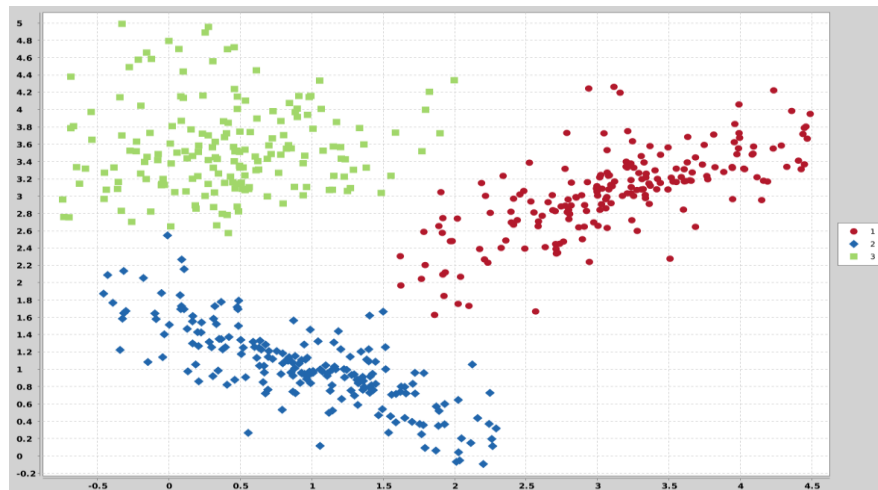


Fig 1.7: K-NN Classification

### 1.5.4 Support Vector Machine

A support vector machine (SVM) is a directed machine learning model that utilizes characterization calculations for two-class order issues. In SVM also calculates the distance from the point like in KNN but here we analyse the perpendicular distance from it.

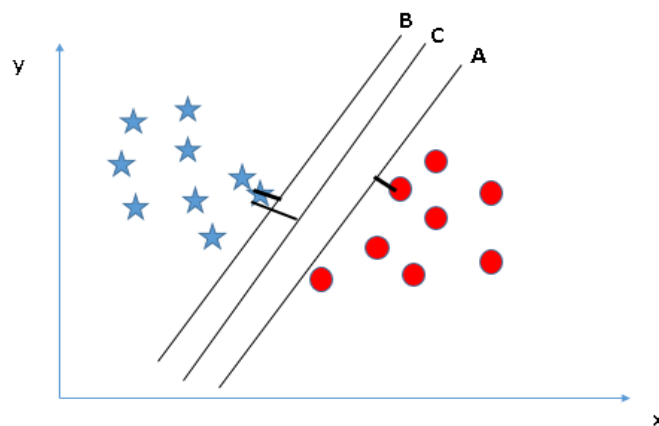


Fig 1.8: Support Vector Machine Classification

## **CHAPTER – 2**

### **LITERATURE SURVEY**

#### **2.1 Prediction Models for Indian Stock Market [4]**

Financial exchange esteem data is made in a very large number and is affected by various factors with every passing minute. Protections trade is an unpredictable and testing system where people will either get money or lose for as far back as they can recollect venture reserves. In this work, an undertaking is made for desire for protections trade design. Two models are manufactured one for step by step estimate and the other one is for month to month desire. Coordinated AI estimations are used to gather the models. As a part of the step by step desire model, chronicled costs are gotten together with sentiments. Up to 70% of precision is seen using controlled AI computations on step by step figure model. Month to month conjecture model endeavours to survey whether there is any comparability between any two months design. Appraisal shows that example of one month is least associated with the example of one more month. There are two basic techniques to anticipate the securities exchange costs. One of the methods that is available for use is the chartist method and the other is the technical theory method. . Proposed method is built on the principle of technical theories. Essential supposition of this hypothesis is history will in general recurrent itself.

##### **2.1.1 Proposed Method**

Past or historical data is taken into account when the model predicts the movement of price  $t_n$ , from  $t_{n-1}, t_{n-2}, \dots, t_1$ , where  $t_n$  is the data for the transaction. Machine learning algorithms are used in order to train the data. Notion from web-based media information and news are separated. Separated assessments later will be coordinated with notable cost to construct expectation model. Clashing assessments has been accounted for by analysts about impact of conclusion on financial exchange. Scarcely any exploration detailed assessment separated from online media has no impact on stock value development though in, they have announced the assumption has either solid or powerless impact on stock value development.

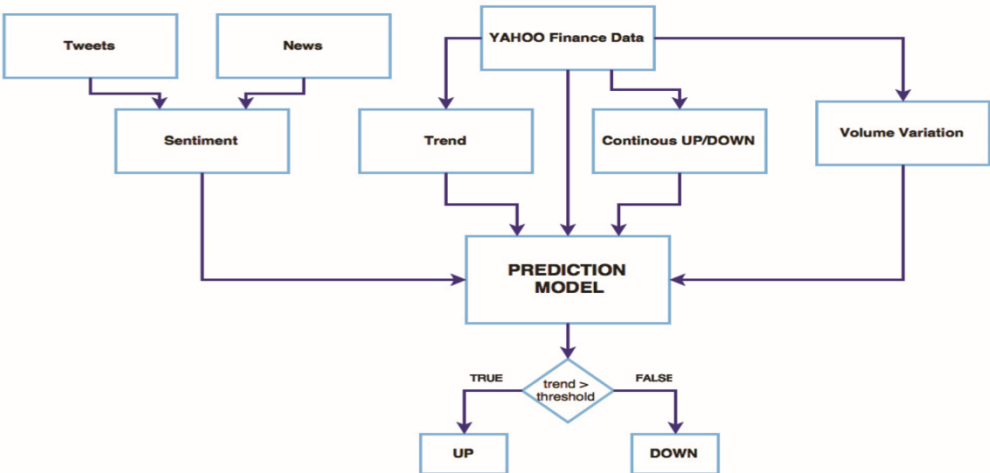
Two unique models have been worked to anticipate financial exchange pattern. First model predicts the securities exchange pattern for the following day (Daily expectation model) by thinking about all accessible information on everyday schedule as info. Second model predicts the financial exchange pattern for the following month (Monthly forecast model) by thinking about

accessible information on month to month premise.

First commitment of the proposed work is that couple of highlights has been derived from the authentic information accessible by utilizing insights. One of the measurable boundaries considered is connection between pattern of a day and volume of stock exchanged around the same time. Volume exchanged element chronicled information will reflect both purchased and sold stocks consistently.

At the point when the pattern is up volume exchanged may show the sold offers, correspondingly when the pattern is down volume exchanged may reflect shares purchased by merchants. This component has been joined with pattern of that day to get whether the volume of stocks is sold or purchased by the dealer. Enormous number of volume exchanged has positive effect if and just if shares are bought by the dealer. Supposition for the offers bought is stock exchanges are more and pattern is down. On the off chance that volume exchanged is more and pattern is up methods, shares are offered to pick up cash. One more measurable boundary is registered by considering past n days example of up/down. These highlights are produced for preparing and testing dataset.

Presently the forecast model is based on preparing dataset. Another commitment of this paper is Monthly expectation model. In this the whole month pattern is processed by thinking about chronicled information. Contribution to the model is given month shrewd. Month ms and year ys expectation depends on year  $m - 1, m - 2, \dots$  of year y. Here the supposition that is pattern of month m in the year y will follow pattern of some extraordinary month in the very year.



**Fig 2.1:** Flow Chart for Daily Prediction Model

### **2.1.2 Conclusion**

In order to make some quick money, it has been seen in the past few years a lot of people are investing in the stock market. But all this comes with a huge risk of losing all your money. Therefore, in order to reduce that risk, a definite predictive model is required. A huge number of models have been put forth to learn about the trends of the market and whether it is going up and down, but an accurate answer to the people's queries is still missing. Hence, in this research paper it is tried to put forward to the people a sort of accurate model to predict those values. Various measures are considered in order to build this model like everyday fluctuations in the market, amount exchanged everyday and mainly the general sentiment of the public which plays a major role in how the company is perceived and how different aspects of the company will turnout.

On the considered dataset, Decision Boosted Tree is performing in a way that is better than Support Vector Machine and Logistic Regression.

## **2.2 Stock Market Prediction Using Machine Learning [5]**

In India, bits of Infosys are recorded on the BSE where it is a bit of the BSE SENSEX and the NSE where it is a NIFTY 50 Constituent. Throughout some stretch of time, the shareholding of its advertisers has steadily decreased, beginning from June 1993 when its offers were first recorded. The sponsors' property diminished further when Infosys transformed into the essential Indian-enlisted association to list Employees Stock Options Schemes and ADRs on NASDAQ on 11 March 1999. The publicist holding tight 31 March 2002 was 28.72% [48] and at 30 June 2017 it dropped to 12.75% as they consistently sold their offers and decreased commitment in powerful organization of the association

Stock Market follows the arbitrary walk, which infers that the best expectation you can have about the upcoming worth is the present worth. Because of the huge fluctuations in the market it is a pretty huge task to develop an accurate model and because of these fluctuations the person who is investing his hard earned money in order to get some little bonuses loses his faith in it. Stock prices change and vary in a blink of an eye and are very dynamic because of the nature of the financial market and because of various other reasons (Previous day's end value, P/E proportion and so forth) and the obscure elements (like Election Results, Rumors and so on). Attempts have been made in order to find a machine learning model that predicts this easily for the people. The main idea when

it comes to projects of research in this field is based on these three main points. The price that is being targeted can change in less than a minute, tomorrow or some time in the coming week or it can be months. The stocks when arranged in set can be less than, to stocks belonging to an industry, to all stocks in general. There can be a number of different sources from where we get our data in order to predict, like it can be from some international news outlet or economic outlet or it can be the general sentiment of the general public towards the company or it can be historical data of the stock prices.

The main target that we want to achieve is to get the future values of the stocks, or to understand the dynamic and volatile nature of the market or to understand the market trend. In the stock market prediction model there is a dummy and a real time prediction. Some set of rules are defined in the dummy prediction and the future values are calculated by using the average price whereas in the real time prediction, it is absolutely compulsory to use the internet and to observe the current prices of different shares.

Computational advances have prompted presentation of AI procedures for the prescient frameworks in monetary business sectors. In this paper we are utilizing a Machine Learning strategy i.e., Support Vector Machine (SVM) to anticipate the securities exchange and we are utilizing Python language for programming.

### **2.2.1 Proposed Method**

The experiments were carried out by Weka and Yale data mining experiments:

The general setup used is as follows:

Part 1:

This step is significant for the download information from the net. We are anticipating the monetary market estimation of any stock. So the offer an incentive up to the end date are download from the website.

Part 2:

In the subsequent stage the information value of any stock that can be changed over into the CSV document (Comma Separate Value) so it will handily stack into the calculation.

Part 3:

In the next stage wherein GUI is open and when we click on the SVM button it will show the window from which we select the stock dataset esteem document.

Part 4:

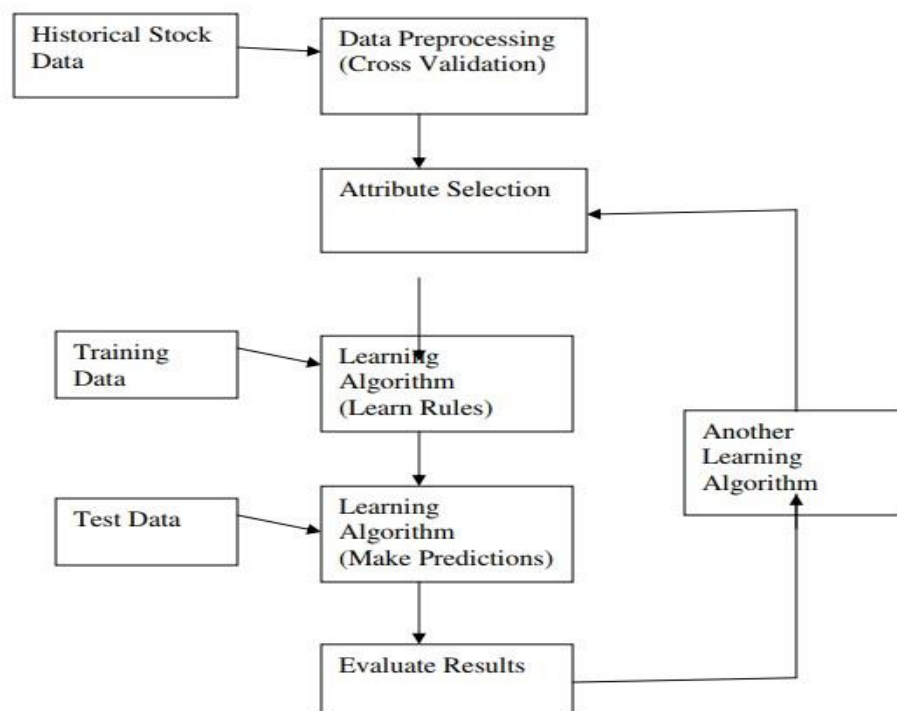
Subsequent to choosing the stock dataset record from the organizer it will show chart Stock prior to planning and stock in the wake of planning.

Part 5:

The subsequent stage calculation determined the  $\log_2 c$  and  $\log_2 g$  esteem for limiting blunder. In this way, it will foresee the diagram for the dataset esteem proficiently.

Part 6:

In definite advance calculation show the anticipated worth diagram of select stock which shows the first worth and anticipated estimation of the stock.



**Fig 2.2:** Learning Environment

## 2.2.2 Conclusion

In the task, we proposed the utilization of the information gathered from various worldwide

monetary business sectors with AI calculations to anticipate the stock file developments. SVM calculation takes a shot at the enormous dataset esteem which is gathered from various worldwide monetary business sectors. Likewise, SVM doesn't give an issue of over fitting. Different AI based models are proposed for foreseeing the every day pattern of Market stocks. Mathematical outcomes recommend the high effectiveness. The viable exchanging models based upon our very much prepared indicator. The model creates higher benefit contrasted with the chose benchmarks.

# CHAPTER – 3

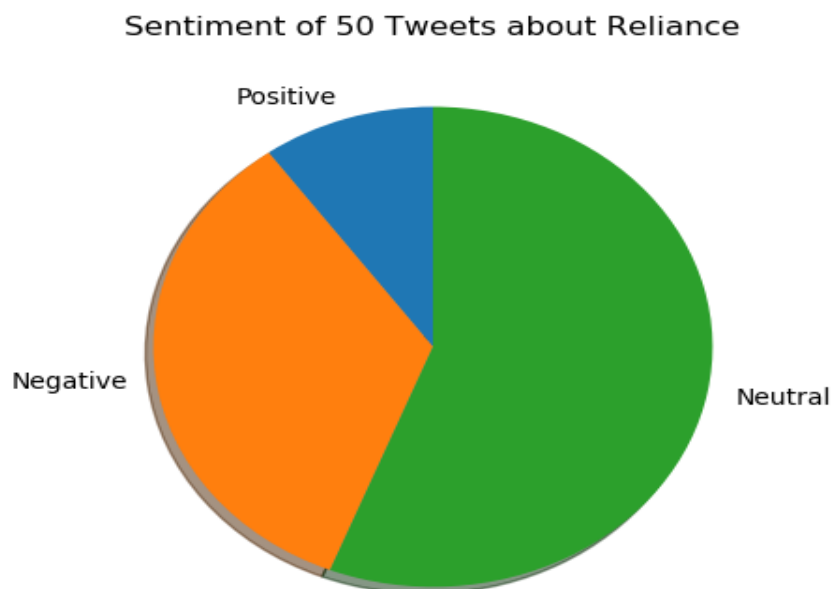
## SENTIMENT ANALYSIS OF DIFFERENT COMPANIES

### 3.1 Reliance

Reliance Industries Limited (RIL) was started by Late Dhirubhai Ambani with his brother in law . Reliance started with manufacturing of women clothes then they expanded to Oil, Telecomm , News , Finance , Energy , Ecommerce etc. Now the company is managed by Mr Mukesh Ambani . People have trusted this company because it's a family business and given exceptional return.

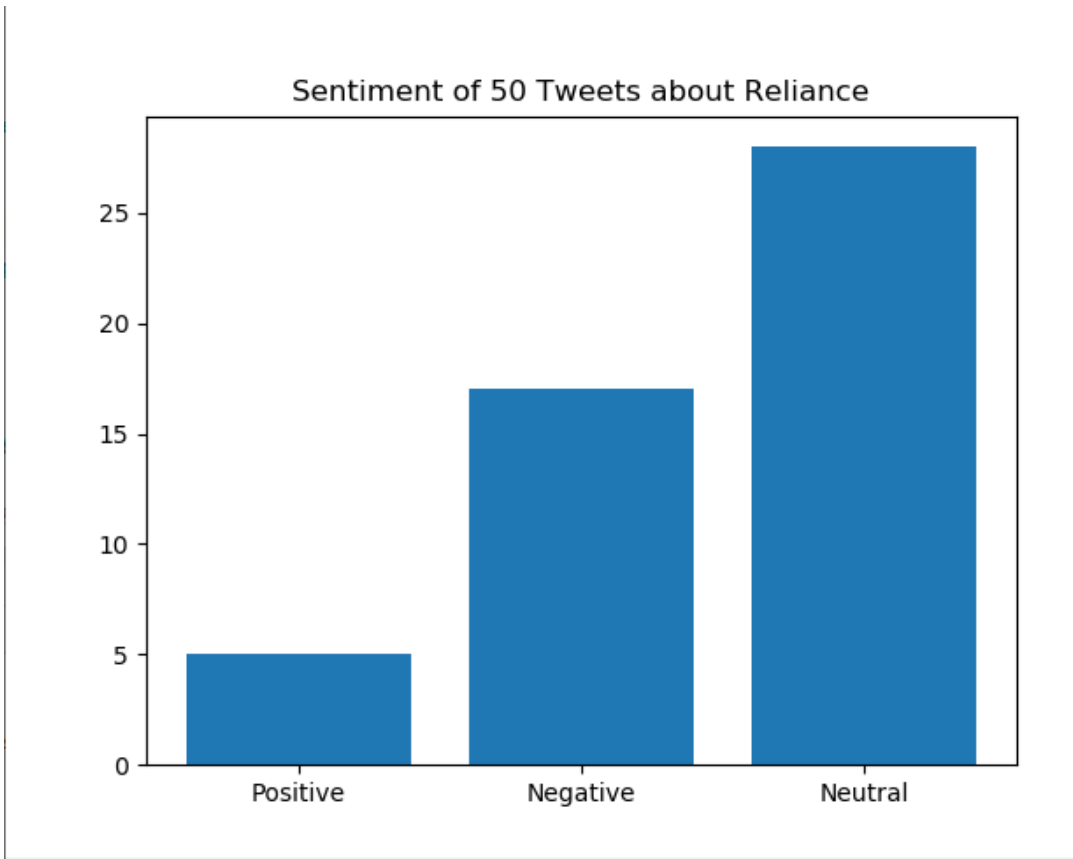
Reliance industries having a main market position and piece of the pie in India which considered as their best strength. Reliance business network isn't simply in India they have business more than five landmasses. Considering the Indian market they have without a doubt, not many contenders to contend. Total Nifty 50 it has 16% of Reliance share just effect on reliance share it can show significant effect on the whole stock market.

#### 3.1.1 Results

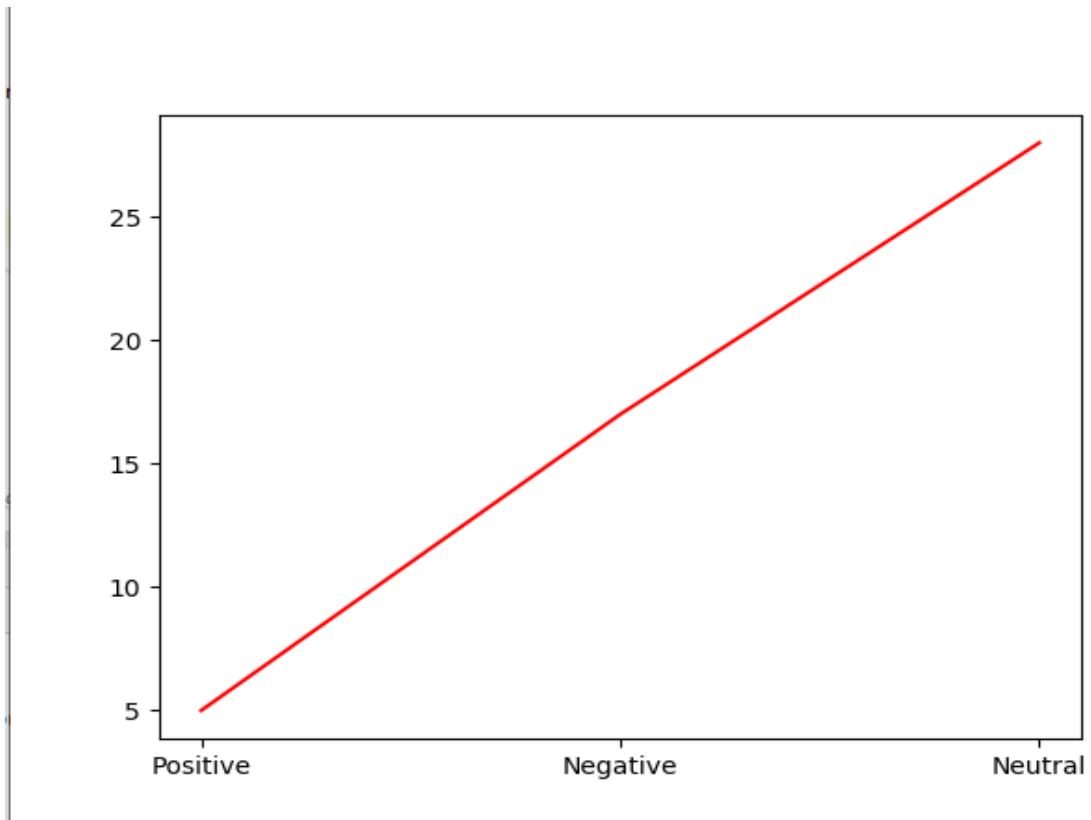


**Fig 3.1:** Pie Chart for Sentiment Analysis on Reliance





**Fig 3.2:** Bar Graph for Sentiment Analysis on Reliance



**Fig 3.3:** Line Graph for Sentiment Analysis on Reliance

### **3.1.2 Conclusion**

After observing the above results we can finally conclude that the general public sentiment, as is derived by using sentiment analysis on 50 twitter users on 29<sup>th</sup> November 2020, is somewhat towards a relatively negative and a hugely towards the neutral side of the spectrum which can be considered as not that great. This type of public sentiment towards the company can have a massively negative impact on the stocks of the company.

Moreover, Dependence Industries on Friday revealed a 15 percent year-on-year drop in united net benefit for the July-September quarter as the COVID-19 pandemic hit its key petrochemicals and oil refining organizations. In the subsequent quarter, its combined net benefit remained at Rs 9,567 crore, contrasted and Rs 11,262 crore in the year prior quarter [7]. This is unquestionably not an incredible sign as the stock costs have likewise gone down for the organization.

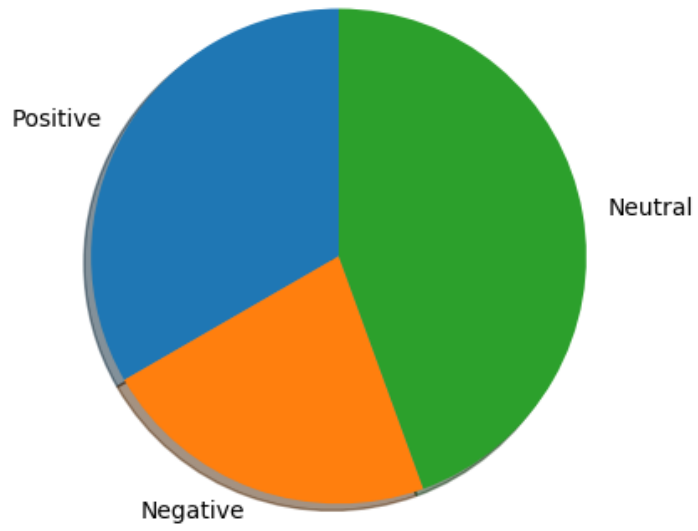
## **3.2 TCS (Tata Consultancy Services)**

Tata Consultancy Services Limited (TCS) is an Indian global data innovation (IT) benefits and counseling organization settled in Mumbai, Maharashtra, India. TCS is the biggest private recruiter in the world . TCS was started by Late FC Kholi who brought IBM in India ,TCS is the reason India is recognised for IT industry . TCS have won many awards for sustainable business model since it fighting against the climate change . TCS has the maximum retaining of employee research have shown employee's are happiest at TCS than other IT firms.

TCS is the second biggest Indian organization by market capitalisation. Tata consultancy associations is correct now arranged among the primary IT associations marks the world over. In 2015, TCS was situated 64th all around in the Forbes World's Most Innovative Companies situating, making it both the most imperative situated IT organizations association and the top Indian association. It is the world's greatest IT organizations supplier. In April 2018, TCS changed into the central Indian IT relationship to reach \$100 billion in market capitalisation, and second Indian affiliation ever (after Reliance Industries accomplished it in 2007 after its market capitalisation remained at ₹6,79,332.81 crore (\$102.6 billion) on the Bombay Stock Exchange.

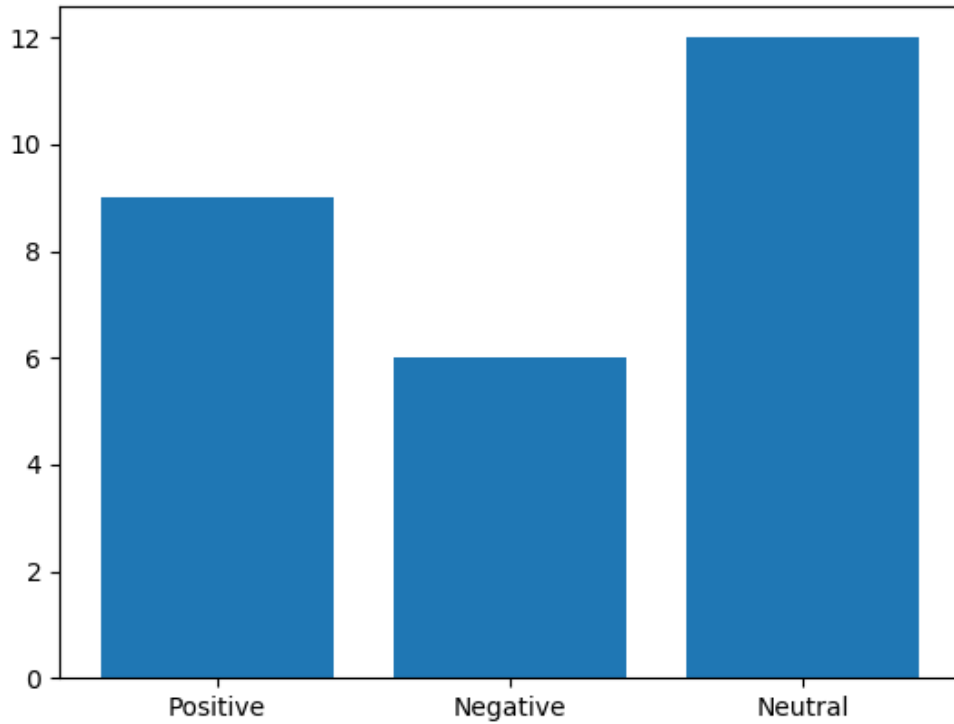
## **Results**

Sentiment of 50 Tweets about



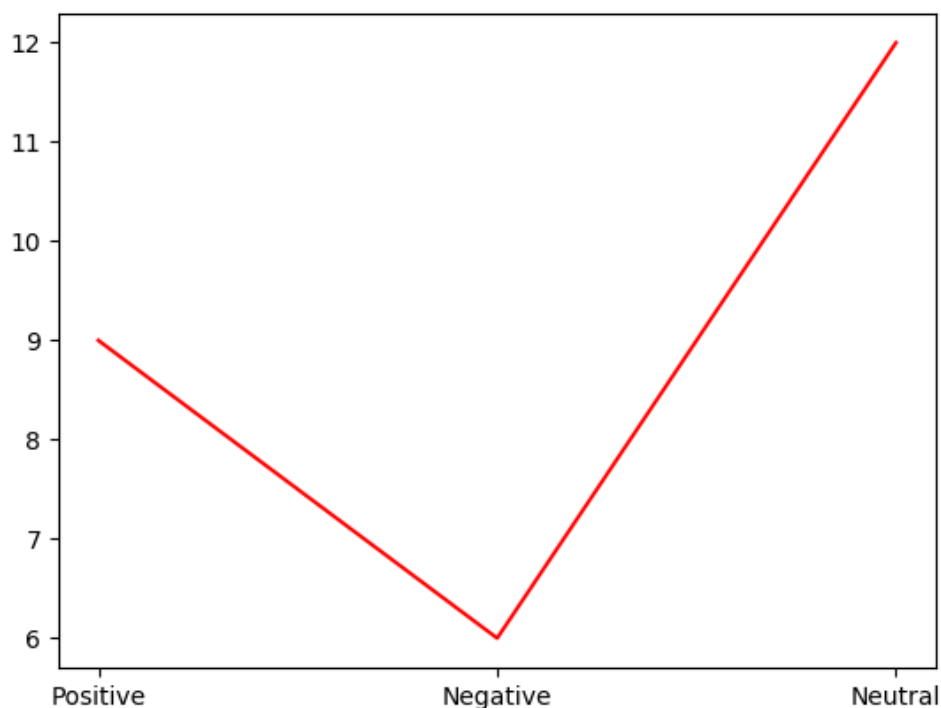
**Fig 3.4:** Pie Chart for Sentiment Analysis on TCS

Sentiment of 50 Tweets about



**Fig 3.5:** Bar Graph for Sentiment Analysis on TCS

17



**Fig 3.6:** Line Graph for Sentiment Analysis on TCS

### 3.2.1 Conclusion

After observing the above results we can finally conclude that the general public sentiment, as is derived by using sentiment analysis on 50 twitter users on 29<sup>th</sup> November 2020, is somewhat towards a relatively positive side of the spectrum which can be considered as great. This type of public sentiment towards the company can have a massively positive impact on the stocks of the company.

The Q1 profits of TCS fell 14 percent as did the profits of many other companies and the Q2 profits declined by 7 percent which is relatively better when we compare that to our previous company, i.e., Reliance. But decline in profits is definitely not a good sign for any company but despite these declines, the company still manages to have a positive public opinion of itself in the market which in the end helps the company's image and its stocks.

### 3.3 Wipro

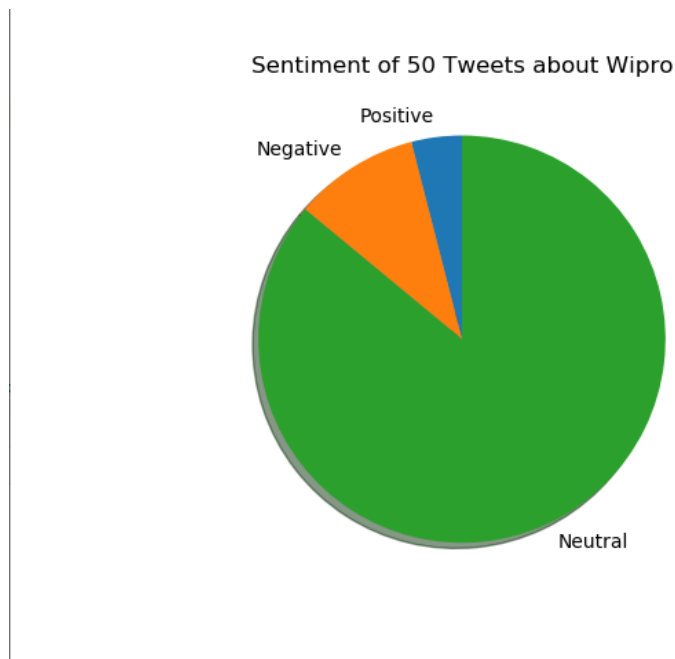
Wipro Limited was started by Mr AzimPremjifather before Independence it was a vegetable oil company then in 1981 expanded into IT industry. It is settled in Bangalore, Karnataka, India.In2013, Wipro isolated its non-IT organizations and framed the exclusive Wipro Enterprises. The organization was joined on 29 December 1945 in Amalner, Maharashtra by Mohamed Premji

as "Western India Palm Refined Oil Limited", later curtailed to "Wipro". It was at first set up as a producer of vegetable and refined oils in Amalner, Maharashtra, British India, under the trademarks of Kisan, Sunflower, and Camel.

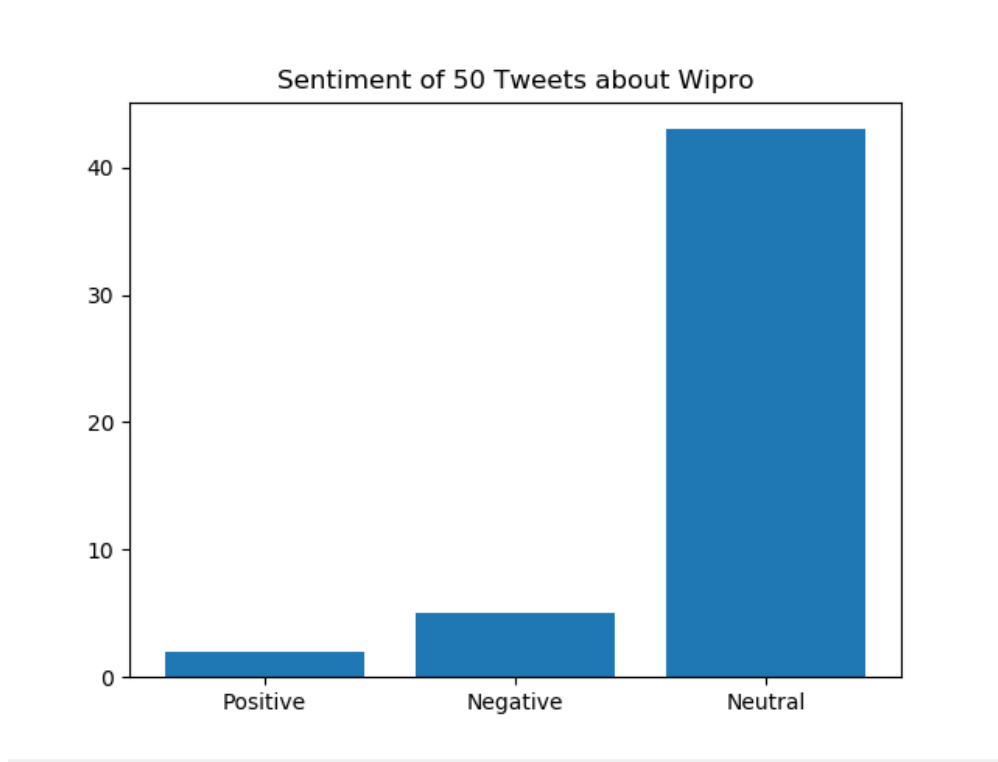
In 1966, after Mohamed Premji's demise, his child AzimPremji took over Wipro as its director at 21 years old.

Wipro's first proposal of stock was in the 1946. Wipro's worth offers are recorded on Bombay Stock Exchange, where it is a constituent of the BSE SENSEX list, and the National Stock Exchange of India where it is a constituent of the S&P CNX Nifty. The American Depository Shares of the affiliation are recorded at the NYSE since October 2000.

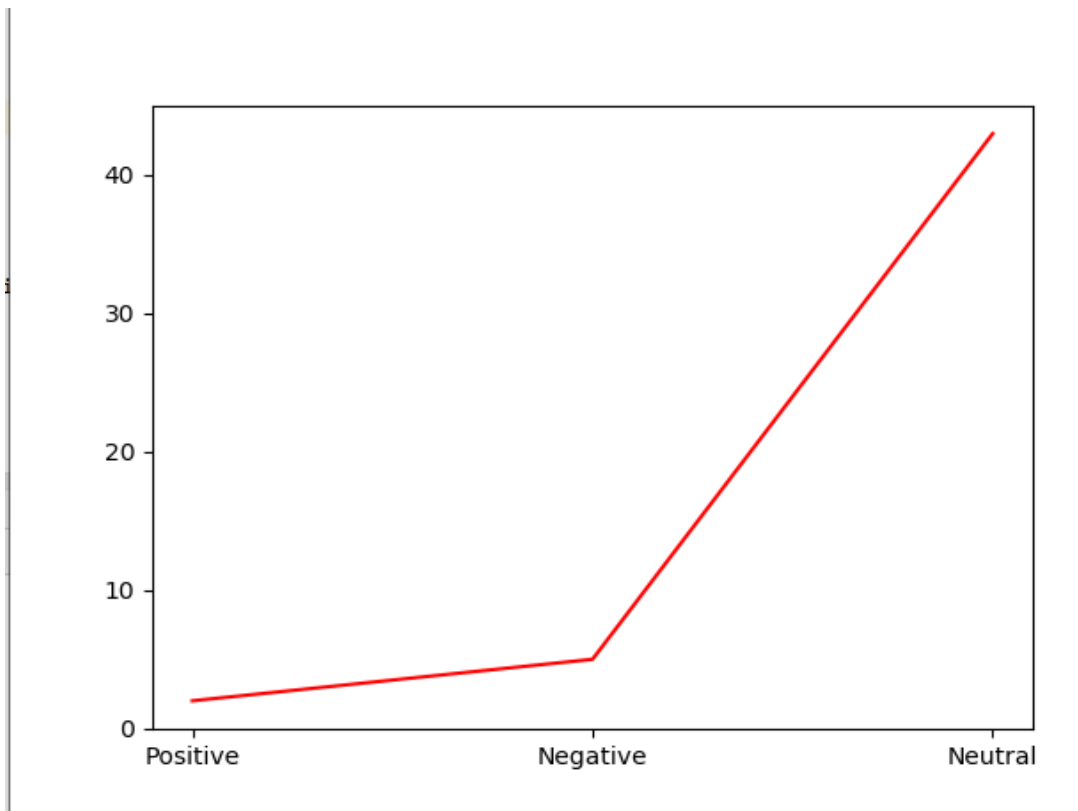
#### Results



**Fig 3.7:** Pie Chart for Sentiment Analysis on Wipro



**Fig 3.8:** Bar Graph for Sentiment Analysis on Wipro



**Fig 3.9:** Line Graph for Sentiment Analysis on Wipro

20

### **3.3.1 Conclusion**

After observing the above results we can finally conclude that the general public sentiment, as is derived by using sentiment analysis on 50 twitter users on 29<sup>th</sup> November 2020, is somewhat towards a relatively neutral side of the spectrum which can be considered as not so great as every company's main focus is always on increasing its profits whereas a neutral sentiment of the public denies that. This type of public sentiment towards the company can have a somewhat negative impact on the stocks of the company but since the last two companies that we worked on were much bigger in comparison to Wipro, therefore, it might not affect it as negatively as you might expect. The neutral sentiment might even work in its favour.

If we look at the net profit that Wipro made this year amid the COVID-19 pandemic, it is quite commendable as it made 1-3 percent of profit this year. This might definitely work in the favour of the company and the sentiment of the public which at the moment is relatively neutral might get more positive as the time passes.

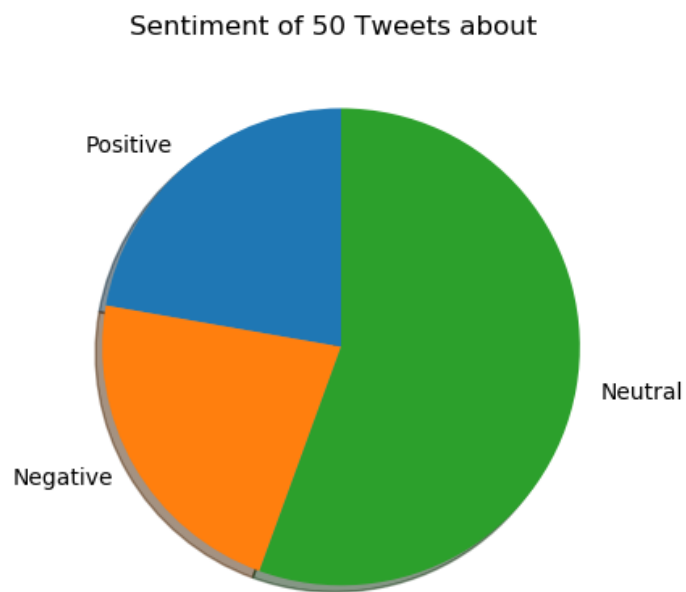
### **3.4 Infosys**

Infosys Limited, is an Indian worldwide association that gives business directing, information advancement and reconsidering organizations. The association is gotten comfortable Bangalore, Karnataka, India. Infosys is the second-greatest Indian IT association after Tata Consultancy Services by 2017 pay figures and the 596th greatest public association on earth reliant on income. On 29 March 2019, its market capitalisation was \$46.52 billion. The FICO appraisal of the association is A- (rating by Standard and Poor's).

Infosys is started by eight people without a computer for seven years it did not made any profit but 1989 after liberalization countries were allowed to come in our country for businesses so it also started generating profit . In 1993 it was listed in stock exchange and made many employees millions. Management of Infosys has been rocky that might be the reason since it is very far away from TCS in market capitalization. Recently Infosys has been acquiring many companies in America showing a Global expansion . Training of Infosys is world renowned which is in Mysore . Infosys is expanding with recent times in many fields such as IOT, Blockchainetc but still the main revenue is generated from customer services.

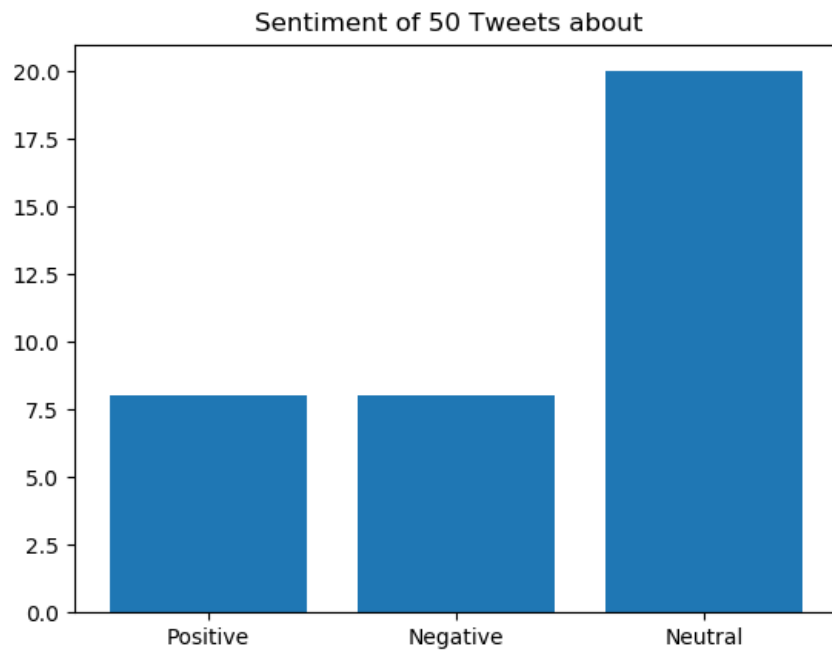
In this pandemic when the MNC like Accenture were laying of the employees Infosys showed the loyalty towards their employees also in this they also gave promotion in this October this shows how strong the company is from inside.

### 3.4.1 Results



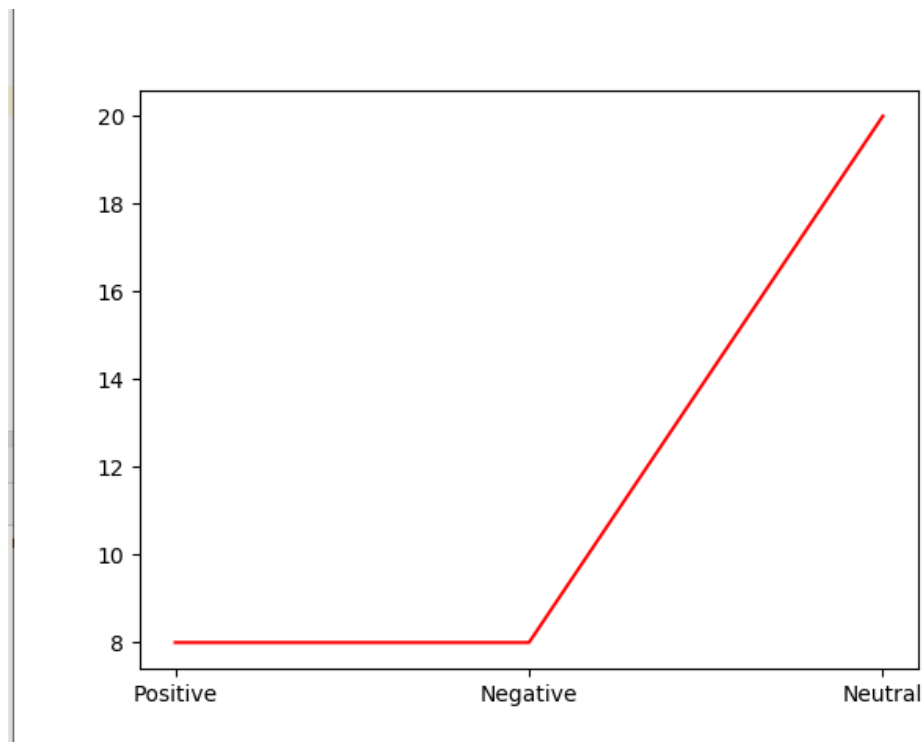
**Fig 3.10:** Pie Chart for Sentiment Analysis on Infosys





**Fig 3.11:** Bar Graph for Sentiment Analysis on Infosys

22



**Fig 3.12:** Line Graph for Sentiment Analysis on Infosys

### 3.4.2 Conclusion

After observing the above results we can finally conclude that the general public sentiment, as is derived by using sentiment analysis on 50 twitter users on 29<sup>th</sup> November 2020, is somewhat towards a relatively neutral side of the spectrum which can be considered as alright. This type of public sentiment towards the company can have a relatively neutral impact on the stocks of the company.

## CHAPTER - 4

# STOCK MARKET PRICE PREDICTION USING TEXTBLOB FOR SENTIMENT ANALYSIS

### 4.1 TextBlob

TextBlob is a library that is made for python (either 2 or 3) to process the vast amounts of data that is produced in the textual format. The superiority of using TextBlob is that it comes up with a relatively easy API, i.e. , an Application Programming Interface to dive into common but multiple natural language processing tasks such as part-of-speech labelling, interpretation, sentiment analysis, classification, translation, and more.

In this particular project we are using this library in order to perform the sentiment analysis on the data that we are retrieving from the twitter about various companies. The sentiment analysis is pretty easy to use in this library by using the inherent sentiment property. This sentiment property set a named tuple of the form SENTIMENT. Polarity basically here means how much is the sentiment of the text leaning towards either of the positive or the negative end of the spectrum or if it is neutral and does not provide any specific sentiment in general. Subjectivity basically refers to how much a tweet offers or how much sentimental data we can get from the text in the tweet. Basically, it uses a Naive Bays Classifier inherently to group the text into either positive or negative polarity. This classifier has been trained on vast amounts of datasets such as movie reviews by various people and then uses the classifier to do the task which it has learned from the movie reviews for example.

### 4.2 Code

#### Importing the necessary libraries

```
In [23]: import tweepy
from textblob import TextBlob
import pandas as pd
import numpy as np
import re
import matplotlib.pyplot as plt
import yfinance as yf
from sklearn.model_selection import train_test_split as tts
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error as mae
from sklearn.metrics import mean_squared_error as mse
from sklearn.metrics import r2_score
from sklearn.preprocessing import LabelEncoder, StandardScaler
```

**Fig 4.1:** Importing necessary libraries

First of all, we import all the necessary libraries that will be required for the full project such as tweepy, textblob, pandas, numpy, etc. All these libraries perform very important steps in the whole project.

## Initializing API Keys and Tokens

```
In [24]: consumerKey = 'eoQrYvBQGiJQNa12ghigGmiIT'
consumerSecret = '1GSGyptLueuboSIzf4g1xxgRSzHYft2Wh7MfvFBO9TbpY0GNZS'
accessToken = '1189221220715155456-RnO4gUXq015ht6bLfOhUgEGHWMo8wB'
accessTokenSecret = 'Bzw3ynkGN13TuG2931ZzrR7WJKFJdi5wf82uiH7raonQi'
```

## Authenticating Keys and Tokens

```
In [25]: authenticate = tweepy.OAuthHandler(consumerKey, consumerSecret)
authenticate.set_access_token(accessToken, accessTokenSecret)
api = tweepy.API(authenticate, wait_on_rate_limit = True)
```

**Fig 4.2:** Initializing and Authenticating Keys and Tokens

Second step that we perform here is to first of all initialize all the API keys and tokens provided by the Twitter Developer site in order to access all the tweets needed and the further, we authenticate those keys and tokens using the tweepy library.

## Getting the Tweets

```
In [26]: posts = tweepy.Cursor(api.search, q='#Infosys', lang='en', tweet_mode='extended').items(500)
df=pd.DataFrame(data = [[tweet.created_at.date(),tweet.full_text]for tweet in posts],columns=['Date','Tweets'])
df
```

Out[26]:

	Date	Tweets
0	2021-03-26	RT @battinaabhisek: #Cipla BUY for the Target...
1	2021-03-26	#Cipla BUY for the Target of Rs 820 \nSL 760 (...)
2	2021-03-26	RT @battinaabhisek: #adaniports Buy for a Targ...
3	2021-03-26	#adaniports Buy for a Target of Rs 750\nWith S...
4	2021-03-26	RT @battinaabhisek: #Bergerpaints need to do...
...	...	...
219	2021-03-19	@MensDayOutIndia @Infosys @Infosys won't care,...
220	2021-03-19	@MariaPramod90 @sourabh_sm @Infosys @zomato @l...
221	2021-03-19	@MariaPramod90 @Infosys @zomato @Infosys won't...
222	2021-03-19	+ve cues (mkt opens gap down): #TCS #Infosys #...
223	2021-03-19	RT @Narryhyd: Congratulations to all my collea...

224 rows x 2 columns

Activate Windows  
Go to PC settings to activate

**Fig 4.3:** Getting the tweets

Next we start to retrieve the tweets that we need by using the property `Cursor` in the library `tweepy`. Here we provide the query and the language for our search and also provide the length of the dataset that we require. Then we convert the data retrieved into a readable tabular format that gives us the date of the tweet in one column and the text in the next one.

## Cleaning the Data

```
In [27]: def cleanText(text):
text = re.sub(r'@[A-Za-z0-9]+', '', text)
text = re.sub(r'#', '', text)
text = re.sub(r':', '', text)
text = re.sub(r'\n', '', text)
text = re.sub(r'_[A-Za-z0-9]+', '', text)
text = re.sub(r'RT[\s]+', '', text)
text = re.sub(r'https?:\\/\s+', '', text)
text = re.sub(r'https?\\/\s+', '', text)
return text

def deEmojiify(txt):
    regex_pattern = re.compile(pattern = "[\u0001F600-\u0001F64F"
                                "\u0001F300-\u0001F5FF"
                                "\u0001F680-\u0001F6FF"
                                "\u0001F1E0-\u0001F1FF"
                                "]+", flags = re.UNICODE)
    return regex_pattern.sub(r'',txt)

df['Tweets'] = df['Tweets'].apply(deEmojiify)
df['Tweets'] = df['Tweets'].apply(cleanText)
```

**Fig 4.4:** Cleaning the data

Next, we clean the data using the Regular Expression library commonly written as 're'. Here we create two functions, one to clean the text of any hashtags, URLs, retweets, etc and the other function is to remove any unwanted emojis. By performing this cleaning process, we get the text in such a format that is much easier for the `TextBlob` library to process and also provides much more accurate results to the sentiment analysis.

## Sentiment Analysis

```
In [28]: def subjectivity(text):  
         return TextBlob(text).sentiment.subjectivity  
def polarity(text):  
         return TextBlob(text).sentiment.polarity  
df['Subjectivity'] = df['Tweets'].apply(subjectivity)  
df['Polarity'] = df['Tweets'].apply(polarity)
```

```
In [29]: def getSentiment(score):  
         if score<0:  
             return 'Negative'  
         elif score == 0:  
             return 'Neutral'  
         else:  
             return 'Positive'  
df['Sentiment'] = df['Polarity'].apply(getSentiment)  
df['Date'] = pd.to_datetime(df['Date'])  
df.drop_duplicates(inplace = True)  
df
```

Out[29]:

	Date	Tweets	Subjectivity	Polarity	Sentiment
0	2021-03-26	Cipla BUY for the Target of Rs 820 SL 760 (Clo...	0.000000	0.000000	Neutral
1	2021-03-26	Cipla BUY for the Target of Rs 820 SL 760 (Clo...	0.000000	0.000000	Neutral
2	2021-03-26	adaniports Buy for a Target of Rs 750With Stop...	0.000000	0.000000	Neutral
3	2021-03-26	adaniports Buy for a Target of Rs 750With Stop...	0.000000	0.000000	Neutral
4	2021-03-26	Bergerpaints need to close above 770 to break...	0.100000	0.000000	Neutral
...	...	...	...	...	...
219	2021-03-19	won't care, they care about their image and...	0.333333	0.250000	Positive
220	2021-03-19	won't care, they care about their image a...	0.333333	0.250000	Positive
221	2021-03-19	won't care, they care about their image an...	0.333333	0.250000	Positive
222	2021-03-19	+ve cues (mkt opens gap down) TCS Infosys HCLT...	0.324495	0.029293	Positive
223	2021-03-19	Congratulations to all my colleagues in the Cl...	0.000000	0.000000	Neutral

Activate Windows  
Go to PC settings to activate

**Fig 4.5:** Sentiment Analysis

Next we perform the main sentiment analysis process by using the TextBlob library and using the sentiment property included in the library. We create two functions namely, subjectivity() and polarity(), and create two news columns of the same name to occupy the values that we get from the sentiment property in TextBlob. Further we make another column to name the sentiment based on the score received and finally we make another dataframe that shows the whole output in a very orderly fashion.

## Clubbing The Data

```
In [32]: lst = list(df['Date'].drop_duplicates())
sen = []
for i in lst:
    x = df[df['Date']== i]
    sen.append(x['Sentiment'].mode()[0])
d = {'Date' : lst, 'Sentiment' : sen}
df1 = pd.DataFrame(data = d, columns = ['Date', 'Sentiment'])
df1
```

```
Out[32]:
```

	Date	Sentiment
0	2021-03-26	Neutral
1	2021-03-25	Neutral
2	2021-03-24	Neutral
3	2021-03-23	Neutral
4	2021-03-22	Positive
5	2021-03-21	Positive
6	2021-03-20	Neutral
7	2021-03-19	Positive

Activate Windows  
Go to PC settings to activate

Fig 4.6: Clubbing the data

Further we club the data together as from twitter we get many tweets for the same day. So we create another dataframe based on the general sentiment on that particular day.

## Getting The Stock Data

```
In [38]: inf = yf.Ticker('INFY')
hist = inf.history(start = '2021-03-19', end = '2021-03-27')
```

```
In [39]: hist.drop(columns = ['Dividends', 'Stock Splits'], inplace=True)
hist.reset_index()
hist
```

```
Out[39]:
```

	Date	Open	High	Low	Close	Volume
0	2021-03-18	18.670000	18.730000	18.350000	18.379999	9617900
1	2021-03-19	18.559999	18.790001	18.389999	18.750000	17281100
2	2021-03-22	18.959999	19.180000	18.799999	19.070000	6428000
3	2021-03-23	18.950001	18.950001	18.700001	18.709999	8851000
4	2021-03-24	18.740000	18.840000	18.580000	18.660000	8412700
5	2021-03-25	18.469999	18.559999	18.250000	18.459999	6647200
6	2021-03-26	18.530001	18.719999	18.299999	18.719999	13422698

```
In [40]: final = pd.merge(hist, df1, on='Date')
final
```

```
Out[40]:
```

	Date	Open	High	Low	Close	Volume	Sentiment
0	2021-03-19	18.559999	18.790001	18.389999	18.750000	17281100	Positive
1	2021-03-22	18.959999	19.180000	18.799999	19.070000	6428000	Positive
2	2021-03-23	18.950001	18.950001	18.700001	18.709999	8851000	Neutral
3	2021-03-24	18.740000	18.840000	18.580000	18.660000	8412700	Neutral
4	2021-03-25	18.469999	18.559999	18.250000	18.459999	6647200	Neutral
5	2021-03-26	18.530001	18.719999	18.299999	18.719999	13422698	Neutral

Activate Windows  
Go to PC settings to activate

Fig 4.7: Getting the stock data

Now we move onto the stock prediction part. So first of all we retrieve the stock data by using the yahoo finance library and get the historical data of the stock that we want. Then we create a dataframe of the same and finally merge the sentiment data and the stock data to create a final dataframe which will be used for the predictions.

```
Training and Testing

In [45]: train, test = tts(final, test_size = 0.3)
train_y = train.Close
test_y = test.Close
del train['Close']
del test['Close']
del train['Date']
del test['Date']
le = LabelEncoder()
train.Sentiment = le.fit_transform(train['Sentiment'])
test.Sentiment = le.fit_transform(test['Sentiment'])
sc = StandardScaler()
train = sc.fit_transform(train)
test = sc.fit_transform(test)

C:\Users\User\anaconda3\lib\site-packages\pandas\core\generic.py:5168: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy
self[name] = value

In [46]: model = LinearRegression()
model.fit(train, train_y)
pred = model.predict(test)
a = r2_score(test_y, pred)
b = mae(test_y, pred)
c = mse(test_y, pred)
print('R2 Score is = ', a)
print('Mean Absolute Error = ', b)
print('Mean Squared Error = ', c)

R2 Score is = -1.4547228718940222
Mean Absolute Error = 0.25
Mean Squared Error = 0.07517605182227471
```

**Fig 4.8:** Training and testing

Finally we perform the training and testing on the data and perform some standard scaling and label encoding. Then we fit and predict the data and see how our model scores on the testing data.

### 4.3 Results

As we can see in the figure above, our model is not performing really well. We get an R2 Score in the negative range which is bad in and of itself but we also get very unrealistic Mean Absolute Error and Mean Squared error. This is because the dataset we have is extremely small and to get some sort of reliable results we need at least 500-1000 rows, but because of the twitter's guidelines, we can only get data for the past 10 days. This restricts us to the use of a very small dataset which is not suitable to perform any sort of predictions.



## CHAPTER – 5

# STOCK PRICE PREDICTION USING NLTK FOR SENTIMENT ANALYSIS

### 5.1 NLTK (Natural Language Tool Kit)

This is a library and a leading platform to perform certain operations and work with data of human language. It has a very straightforward interface with many corpora and lexical resources which provide various functionalities such as classification, tokenization, stemming, etc. Some of these processes are used in the sentiment analysis property of the Natural Language Tool Kit. If you want to play with natural language then there cannot be a better platform or library to begin doing it other than Natural Language Tool Kit.

Now, to perform the sentiment analysis on any given text, first of all, the text must be clean. After that the property inside the NLTK library, i.e., `sentiment.vader`, is used. VADER stands for Valence Aware Dictionary for Sentiment Reasoning. Here we have the `SentimentIntensityAnalyzer` that calculates the polarity score. The polarity score here is a dictionary of 4 values namely, `pos` for positive, `neg` for negative, `neu` for neutral and `comp` for compound. Compound value here is basically the total summed up score that we are going to require.

### 5.2 Code

#### Importing Necessary Libraries

```
In [31]: import pandas as pd
import numpy as np
import re
import matplotlib.pyplot as plt
import yfinance as yf
from sklearn.model_selection import train_test_split as tts
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.metrics import mean_absolute_error as mae
from sklearn.metrics import mean_squared_error as mse
from sklearn.metrics import r2_score
from sklearn.preprocessing import LabelEncoder, StandardScaler
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

**Fig 5.1:** Importing Necessary Libraries

Again, we import all the necessary libraries that will be required for the full project such as `re`, `yfinance`, `nltk`, `pandas`, `numpy`, etc. All these libraries perform very important steps in the whole project.

## Reading The Scraped Data

```
In [32]: df = pd.read_csv(r'D:\Study Material\Project\2021.csv')
df
```

Out[32]:

		g_14bl	PT3
0	Robust regulatory framework needed to deal wit...	5.18 pm   09 May 2021	Source:
1	TCS Consolidated March 2021 Net Sales at Rs 43...	8.43 am   07 May 2021	Source:
2	Buy Tata Consultancy Services; target of Rs 37...	4.26 pm   16 Apr 2021	Source:
3	Buy Tata Consultancy Services; target of Rs 36...	1.53 pm   15 Apr 2021	Source:
4	Trade Spotlight: What should investors do with...	8.42 am   15 Apr 2021	Source:
...	...	...	...
671	TCS Q3 PAT may dip 5.1% to Rs 6252.3 cr. Motil...	6.19 pm   11 Jan 2017	Source:
672	TCS Q3 net seen down 1.5%, currency headwind m...	3.14 pm   11 Jan 2017	Source:
673	Boosters: 10 stocks that Deutsche Bank is bull...	1.53 pm   10 Jan 2017	Source:
674	Buy, sell, hold: 21 large & midcap stocks to b...	10.02 am   10 Jan 2017	Source:
675	Here are Sanjiv Bhasin's top trading ideas	9.28 am   10 Jan 2017	Source:

676 rows x 2 columns

Activate Windows

**Fig 5.2:** Reading the Scraped Data

For this part of the project we had to increase our dataset substantially, which was just comprised of 10 rows of data from the twitter. But now we performed some web scraping on the financial news site called as moneycontrol.com and scraped the headlines including news about TCS for the past 3 years dating back to 2017. After scraping the data we saved it into CSV format and then further cleaning was performed.

## Cleaning Data

```
In [33]: def get_date(text):
    date = re.search(r'\d\d[s[A-Za-z]+]\s\d\d\d\d', text)
    if date:
        return date.group(0)
    return ""
df['Date'] = df['PT3'].apply(get_date)

df.rename(columns = {'g_14bl':'Headlines'}, inplace=True)

df['Date'] = pd.to_datetime(df['Date'])

df.drop_duplicates(inplace=True)

df = df.reset_index()

del df['PT3']
del df['index']
```

**Fig 5.3:** Cleaning the Data

Data cleaning in this case was pretty straight forward as the headlines only contained proper text, not like the tweets which included hashtags, emojis, retweets, etc.

We just needed to extract the date from the column which included the time, date and source for the news and we did so by performing a simple regular expression search. After that we just renamed the columns and converted the Date column to datetime format. Further we dropped some duplicate rows and deleted the unnecessary columns.

## Sentiment Analysis Using NLTK

```
In [34]: def polarity(text):  
         return SentimentIntensityAnalyzer().polarity_scores(text)  
df['Polarity'] = df['Headlines'].apply(polarity)  
df
```

```
In [35]: for i in range(0,653):  
         df['Polarity'][i] = df['Polarity'][i]['compound']
```

```
In [36]: def getSentiment(score):  
         if score<0:  
             return 'Negative'  
         elif score>0:  
             return 'Positive'  
         else:  
             return 'Neutral'  
df['Sentiment'] = df['Polarity'].apply(getSentiment)  
  
df = df.drop(columns = ['Headlines', 'Polarity'])  
df
```

**Fig 5.4:** Sentiment Analysis Using NLTK

Now we start performing the sentiment analysis on our dataset which is very much similar to the way we performed sentiment analysis by using TextBlob. But the difference is just that we get a much better polarity score when we use the NLTK library. So by running the first line we get the polarity scores in the form of dictionaries from which we extract the compound values by running the second line and then further on, we make another column named sentiment based on the score that we get in our polarity column. Finally, we drop the headlines and polarity column as they are not needed anymore.

## Clubbing Data

```
In [37]: lst = list(df['Date'].drop_duplicates())
sen = []
for i in lst:
    x = df[df['Date']== i]
    sen.append(x['Sentiment'].mode()[0])
d = {'Date' : lst, 'Sentiment' : sen}
df = pd.DataFrame(data = d, columns = ['Date', 'Sentiment'])
df
```

```
Out[37]:
```

	Date	Sentiment
0	2021-05-09	Positive
1	2021-05-07	Neutral
2	2021-04-16	Neutral
3	2021-04-15	Neutral
4	2021-04-14	Neutral
...	...	...
351	2017-01-14	Neutral
352	2017-01-13	Neutral
353	2017-01-12	Neutral
354	2017-01-11	Neutral
355	2017-01-10	Positive

Activate Windows  
Go to PC settings to activate Windows.

**Fig 5.5:** Clubbing the Data

Further we club the data together as we have got multiple news from one particular day here. So we create another dataframe based on the general sentiment on that particular day by performing a mode function on loop.

## Retrieving Historical Stock Data

```
In [40]: inf = yf.Ticker('TCS')
hist = inf.history(start = '2017-01-10', end = '2021-05-09')

hist.drop(columns = ['Dividends', 'Stock Splits'], inplace=True)
hist.reset_index()
hist
```

```
Out[40]:
```

Date	Open	High	Low	Close	Volume
2017-01-09	5.468478	5.505870	5.290870	5.356305	262500
2017-01-10	5.421739	5.636739	5.421739	5.496522	249100
2017-01-11	5.524565	5.524565	5.290869	5.384348	179800
2017-01-12	5.403044	5.403044	5.141304	5.281522	294700
2017-01-13	5.290869	5.384348	5.188044	5.225435	135100
...	...	...	...	...	...
2021-05-03	14.260000	14.480000	13.310000	13.480000	1838400
2021-05-04	13.310000	13.450000	12.760000	13.230000	877200
2021-05-05	13.400000	13.400000	12.460000	12.640000	618400
2021-05-06	12.810000	13.300000	12.680000	13.230000	813300
2021-05-07	13.270000	13.690000	13.160000	13.620000	638100

1090 rows x 5 columns

```
In [41]: final = pd.merge(hist, df, on='Date')
final
```

Activate Windows  
Go to PC settings to activate Windows.

**Fig 5.6:** Retrieving Historical Stock Data

Now we move onto the stock prediction part. So first of all we retrieve the stock data by using the yahoo finance library and get the historical data of the stock that we want. Then we create a dataframe of the same and finally merge the sentiment data and the stock data to create a final dataframe which will be used for the predictions.

## Training and Testing

```
In [42]: train, test = tts(final, test_size = 0.3)
train_y = train.Close
test_y = test.Close
del train['Close']
del test['Close']
del train['Date']
del test['Date']
le = LabelEncoder()
train.Sentiment = le.fit_transform(train['Sentiment'])
test.Sentiment = le.fit_transform(test['Sentiment'])
sc = StandardScaler()
train = sc.fit_transform(train)
test = sc.fit_transform(test)
```

C:\Users\User\anaconda3\lib\site-packages\pandas\core\generic.py:5168: SettingWithCopyWarning:  
A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/indexing.html#returning-a-view-versus-a-copy](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)  
self[name] = value

**Fig 5.7:** Training and Testing

Now we just split our data into training and testing sets and just perform some basic label encoding functions to change our strings to numbers and use some standard scaling in order to scale the data to get better results. Now we apply various regression models.

## Linear Regression

```
In [43]: LR = LinearRegression()
LR.fit(train, train_y)
pred = LR.predict(test)
a = r2_score(test_y, pred)
b = mae(test_y, pred)
c = mse(test_y, pred)
print('R2 Score is = ', a)
print('Mean Absolute Error = ', b)
print('Mean Squared Error = ', c)
```

R2 Score is = 0.9962824636589753  
Mean Absolute Error = 0.1643636192879737  
Mean Squared Error = 0.03304919533060516

**Fig 5.8:** Linear Regression

First, we apply a simple linear regression model to our training and testing dataset. As we can see

we are getting a very high R2 Score, which can either mean that our model is perfect in every sense which is quite impossible or the other possible explanation is that our model is over fitting. But either way we are getting pretty satisfactory results.

## Ridge Regression

```
In [44]: RR = Ridge()
RR.fit(train, train_y)
pred = RR.predict(test)
x = r2_score(test_y, pred)
y = mae(test_y, pred)
z = mse(test_y, pred)
print('R2 Score is = ', x)
print('Mean Absolute Error = ', y)
print('Mean Squared Error = ', z)

R2 Score is = 0.995098166940638
Mean Absolute Error = 0.18166428671377158
Mean Squared Error = 0.04357768785448367
```

**Fig 5.9:** Ridge Regression

Second model that we have here is the Ridge Regression model which is giving us not that different result. We can get better result, maybe, by tweaking the value of alpha which by default is set to 1. But as of now we can see that this model too is over fitting our data.

## Lasso Regression

```
In [45]: LaR = Lasso()
LaR.fit(train, train_y)
pred = LaR.predict(test)
i = r2_score(test_y, pred)
j = mae(test_y, pred)
k = mse(test_y, pred)
print('R2 Score is = ', i)
print('Mean Absolute Error = ', j)
print('Mean Squared Error = ', k)

R2 Score is = 0.8625489632014918
Mean Absolute Error = 0.8882773039485881
Mean Squared Error = 1.221950707897048
```

**Fig 5.10:** Lasso Regression

Lastly, we have Lasso Regression model which gives us some promising result. Here the R2 Score is just in the right spot, which is, not over fitting.

## 5.3 Results

As we can see that there has been a substantial difference in the values of our performance measures from the twitter dataset code and we can definitely attribute all of this to, first of all, to the

large dataset that has been gathered and hence it provides much more data for our models to train on and, secondly, to the Natural Language Tool Kit, which performs an exceptional job at detecting the sentiment of the given text. We are getting an R2 score of 0.99 which is pretty unrealistic and shows that our model is over fitting but the Lasso Regression model gives us a very genuine R2 score of 0.86 which can be said as a good fit for our data.

## **CHAPTER – 6**

### **CONCLUSION**

Nowadays, Data Science, Machine Learning and Artificial Intelligence are gaining a lot of traction and are considered to be the hottest subjects for the coming decades. A huge amount of research work is going on in these fields and a lot of it has got into the markets already such as the recommendation systems.

Just like the recommendation systems that take into account the reviews written by a person, and the use of natural language processing in this process, we have also used natural language processing in a new light, i.e., for the financial market.

In the first code we use TextBlob library for the sentiment analysis on the data collected from the twitter and then predict the future trend of the stock prices and how they will fluctuate.

In the second code we use a much larger and better dataset of news headlines over the years and also use a much larger library for language processing known as Natural Language Tool Kit. Further we use three different models to fit on our data and finally, we get some reliable results.



## REFERENCES

- [1] YazhiGao, WengeRong, YikangShen, Zhang Xiong, “Convolutional Neural Network based sentiment analysis using Adaboost combination”, International Joint Conference on Neural Networks (IJCNN), 2016, Pg 1333-1338.
- [2] Y. Kim and S. Myaeng. “Opinion analysis based on lexical clues and their expansion” . In Proceedings of 6th NTCIR Evaluation Workshop, 2007.
- [3] <https://twitter.com/JumptuitNow/status/1323447352661856256?s=20>
- [4] AparnaNayak, M.M. ManoharaPai, Radhika M. Pai, “Prediction Models for Indian Stock Market”, Twelfth International Multi-Conference on Information Processing,2016, Pg 441-449.
- [5] V KranthiSai Reddy, “Stock Market Prediction Using Machine Learning”, International Research Journal of Engineering and Technology(IRJET), Vol. 5, Issue 10.
- [6] "Top companies in India by Net Profit". Moneycontrol.com.
- [7] <https://www.theweek.in/news/biz-tech/2020/10/30/covid-19-impact-reliance-q2-net-profit-falls-15-from-year-ago.html>