Machine Learning for Long Term Stock Price Prediction

"submitted towards partial fulfilment of the criteria for award of PGPBA by GLIM"

Submitted by:	Sahil Makkar, Pavan Kumar Thatha, Ashwini Murthy, Deepti Ramani, Showbhik, Utkarsh Kulshrestha
Group No:	4
Batch:	PGPBA-BI 2017, Section A
Project Mentor:	Neelesh Singh

GREAT LAKES INSITUTE OF MANAGEMENT, BANGALORE

Acknowledgements

We wish to place on record our deep appreciation for the guidance and help provided to us by our Mentor Mr. Neelesh Singh from Great Lakes Institute of Management for guiding us in completing this project on time.

We would also like to place on record our appreciation for the guidance provided by Dr. P K Vishwanathan for giving us valuable feedback and being a source of inspiration in helping us to work on this project.

Last but not the least we wish to thank Project Office, Bangalore for their support.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: 29 November 2017

Place: Bangalore

Sahil Makkar

Pavan Thatha

Ashwini Murthy

Deepthi Ramani

Utkarsh Kulshrestha

Showbhik

GREAT LAKES INSTITUTE OF MANAGEMENT

Postgraduate Program Business Analytics and Business Intelligence

Certificate

This is to certify that the PGPBA Capstone Project titled "***Machine Learning for Long Term Stock Price Prediction***" is a bonafide record of the Project work carried out by *Sahil Makkar, Utkarsh Kulshrestha, Ashwini Murthy, Deepthi Ramani, Pavan Thatha and Showbhik* in fulfilment of requirements for the award of PGPBA-BI in Great Lakes Institute of Management.

November 2017

Neelesh Singh	**Dr. P K Vishwanathan**
Project Mentor	Program Director

Date: November 2017

Place: Bangalore

Abstract

The Stock market process is full of uncertainty and is affected by many factors. Hence the Stock market prediction is one of the important exertions in finance and business. There are two types of analysis possible for prediction, technical and fundamental. In this paper both technical and fundamental analysis are considered. Technical analysis is done using historical data of stock prices by applying machine learning and fundamental analysis is done using financial performance indicators of the particular business. These days, social media provides expanded opportunity for technical analysis. Social media data has high impact today than ever, it can aide in predicting the trend of the stock market. Assuming that news articles have impact on stock market, this is an attempt to study relationship between news and stock trend. The method involves collecting news and social media data and extracting sentiments expressed by individual. Then the correlation between the sentiments and the stock values is analysed. The learned model can then be used to make future predictions about stock values. It can be shown that this method can predict the sentiment and the stock performance and its recent news and social data are also closely correlated.

- Techniques: Predictive modelling and Machine Learning
- Tools: R, Python, Tableau

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The stock market is essentially a non-linear, nonparametric system that is extremely hard to model with any reasonable accuracy. Investors have been trying to find a way to predict stock prices and to find the right stocks and right timing to buy or sell. To achieve those objectives, and per some research used the techniques of fundamental analysis, where trading rules are developed based on the information associated with macroeconomics, industry, and company. The authors of some of the research said that fundamental analysis assumes that the price of a stock depends on its intrinsic value and expected return on investment. Analysing the company's operations and the market in which the company is operating can do this. Consequently, the stock price can be predicted reasonably well. Most people believe that fundamental analysis is a good method only on a long-term basis. However, for short- and medium-term speculations, fundamental analysis is generally not suitable. Some other research used the techniques of technical analysis, in which trading rules were developed based on the historical data of stock trading price and volume. Technical analysis as illustrated in and refers to the various methods that aim to predict future price movements using past stock prices and volume information. It assumes that history repeats itself and that future market directions can be determined by examining historical price data. Thus, it is assumed that price trends and patterns exist that can be identified and utilized for profit. Most of the techniques used in technical analysis are highly subjective in nature and have been shown not to be statistically valid.

However, now data mining techniques and artificial intelligence techniques like decision trees, rough set approach, and artificial neural networks have been applied to this area. Artificial neural networks can be trained by using an appropriate learning algorithm. Following the assumption of technical analysis that patterns exist in price data, it is possible in principle to use data mining techniques to discover these patterns in an automated manner. Once these patterns have been discovered, future prices can be predicted.

In this project, we will Introduce a framework in which we integrate user predictions into the current machine learning algorithm using public historical data & text mining to improve our results. Following the assumption of technical analysis that patterns exist in price data, it is possible in principle to use data mining techniques to discover these patterns in an automated manner. Once these patterns have been discovered, future prices can be predicted.

Objective:

The main objective of our project is to find out how to predict the future market movement using sentiment data which reflects the reaction of news articles and updates from popular news channels. Another objective is to use machine learning to stabilize the algorithm and see if it can lower the risk factors involvement.

Fundamental Analysis:

Fundamental Analysis is a method of evaluating a security in an attempt to measure its intrinsic value, by examining related economic, financial and other qualitative and quantitative factors. Fundamental analysts study anything that can affect the security's value, including macroeconomic factors such as the overall economy and industry conditions and microeconomic factors such as financial conditions and company management. The end goal of fundamental analysis is to produce a quantitative value that an investor can compare with a security's current price, thus indicating whether security is undervalued or overvalued.

Fundamental analysis involves the process of:

- Looking at the basics or fundamental quantitative and financial environment of a business.
- Examining key ratios to determine financial health
- Estimating the value of the stock

For stocks and equity instruments, this method uses revenues, earnings, future growth, ROE, profit margins and other data to determine a company's underlying value and potential for future growth. In terms of stocks, fundamental analysis focuses on the financial statements of the company being evaluated.

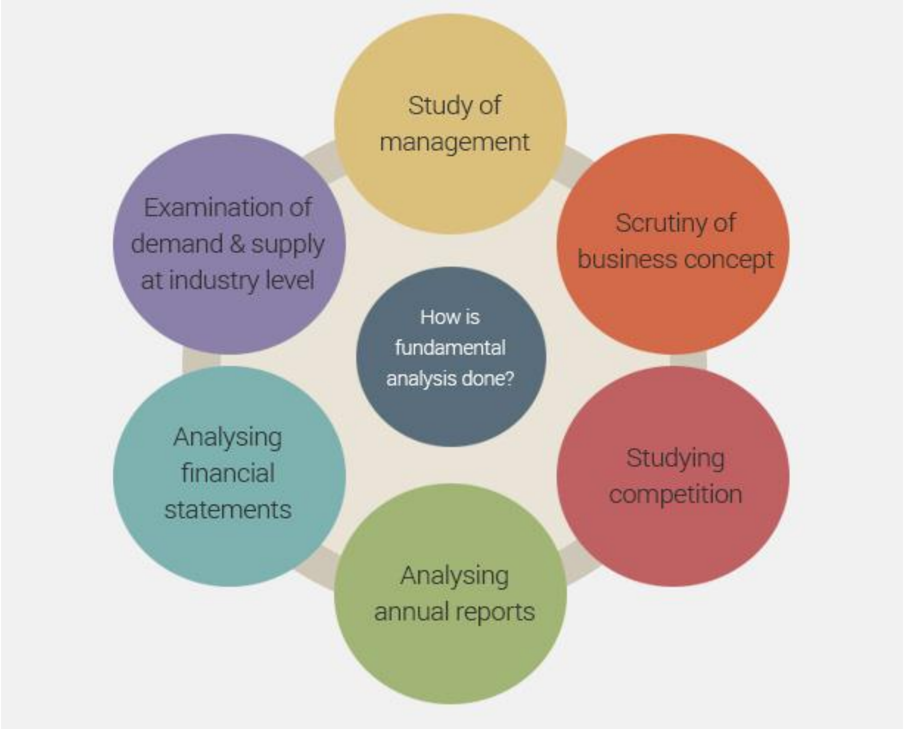

Study of management
Scrutiny of business concept
Studying competition
Analysing annual reports
Analysing financial statements
Examination of demand & supply at industry level
How is fundamental analysis done?

Fig 1: Fundamental Approach

The goal of analysing a company's fundamentals is to find a stock's intrinsic value, a term which gives what a stock in really worth – as opposed to a value at which it is being traded. If the intrinsic value is more than the current share price, our analysis is showing that the stock is worth more than its price and that it makes sense to buy this stock. In terms of stocks, fundamental analysis focuses on the financial statements of the company being evaluated.

Importance of Stock Price Prediction

In the finance field, stock market and its trends are extremely volatile in nature. It attracts researchers to capture the volatility and predicting its next moves. Investors and market analysts study the market behaviour and plan their buy or sell strategies accordingly. As stock market produces large amount of data every day, it is very difficult for an individual to consider all the current and past information for predicting future trend of a stock. Mainly there are two methods for forecasting market trends. One is Technical analysis and other is Fundamental analysis. Technical analysis considers past price and volume to predict the future trend whereas Fundamental Analysis on the other hand, Fundamental analysis of a business involves analysing its financial data to get some insights. The efficacy of both technical and fundamental analysis is disputed by the efficient-market hypothesis which states that stock market prices are essentially unpredictable. This research follows the Fundamental analysis technique to discover future trend of a stock by considering news articles about a company as prime information and tries to classify news as good (positive) and bad (negative). If the news sentiment is positive, there are more chances that the stock price will go up and if the news sentiment is negative, then stock price may go down. This research is an attempt to build a model that predicts news polarity which may affect changes in stock trends. In other words, check the impact of news articles on stock prices. We are using supervised machine learning as classification and other text mining techniques to check news polarity. And, be able to classify unknown news, which is not used to build a classifier. We have taken historical, up to last 5 years' data for analysis.

Important Terminologies in Stock Market

Beta: A measurement of the relation between the price of the stock and the movement of the whole market. If stock XYZ has a beta of 1.5, that means for every 1-point movement in the market, stock XYZ moves 1.5 points and vice versa.

Broker: A person who buys and sells the investment for you in exchange for a fee (a commission).

Dividend: Portion of the company's earnings that is paid to the shareholders, or people what own that company's stock, on a quarterly or annual basis.

Exchange: An exchange is a place in which different investments are traded. For eg. The NSE, BSE in India, NYSE and NASDAQ in U.S

Financial Statements: Formal record of the financial activities and position of a business, person or other entity. Relevant financial information is presented in a structured manner and a form easy to understand.

Index: Index is a benchmark which is used as a reference marker for traders and portfolio managers. A 10% may sound good, but if the market index returned 12%, then you didn't do very well since you could have just invested in an index fund and saved time by not trading frequently.

Intrinsic Value: Refers to the value of the company, stock or product determined through fundamental analysis without reference to its market value.

Margin: A margin account lets a person borrow money (take out a loan essentially) from a broker to purchase an investment. The difference between the amount of the loan, and the price of the securities, is called the margin.

Moving Average: A stock's average price-per-share during a specific period of time. Some time frames are 50 and 200 day moving averages.

Portfolio: A collection of investments owned by an investor. You can have as little as one stock in a portfolio to an infinite amount of stocks.

Quote: Information on a stock's latest trading price. This

Sector: A group of stocks that are in the same business. An example would be the "Technology" sector including companies like Apple and Microsoft.

Stock Symbol: A one-character to three-character, alphabetic root symbol, which represents a publicly traded company on a stock exchange. Apple’s stock symbol is AAPL.

Volatility: This refers to the price movements of a stock or the stock market as a whole. Highly volatile stocks are ones with extreme daily up and down movements and wide intraday trading ranges.

Volume: The number of shares of stock traded during a particular time period, normally measured in average daily trading volume.

Yield: This usually refers to the measure of the return on an investment that is received from the payment of a dividend. This is determined by dividing the annual dividend amount by the price paid for the stock.

Chapter 2

Financial Ratios

The main objective of a financial a report is to provide information on a company's performance to the internal and external users to take decisions. From the perspective of information economics, accounting and financial reporting play a vital role in an efficient capital market. One of the most common ways of assessing the relative values of stocks among practitioners is to compare the numbers listed in financial statements by using financial ratios. The main advantage of using financial ratios instead of amounts from the income statement is that they are independent of the size of the company. The comparison of financial ratios is used to assess companies' financial condition, operations and attractiveness as an investment.

Financial ratios defined as the numerical value created from two or more values taken from a company's financial statements i.e. its balance sheet, income statement or statement of cash flow. Typically, financial ratios are presented as a quantified metric in the form of a percentage, multiple or a ratio which aims to evaluate the financial, operational performance and competitiveness of a company.

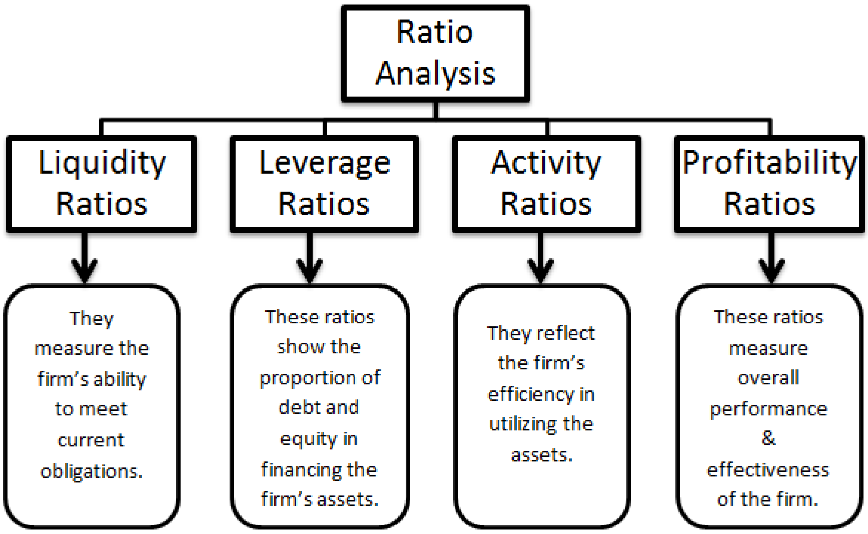

Ratio Analysis
Liquidity Ratios
Leverage Ratios
Activity Ratios
Profitability Ratios
They measure the firm's ability to meet current obligations.
These ratios show the proportion of debt and equity in financing the firm's assets.
They reflect the firm's efficiency in utilizing the assets.
These ratios measure overall performance & effectiveness of the firm.

Fig 2: Financial Ratio Analysis

Liquidity Ratios:

The liquidity focuses on a firm's ability (current assets and current liabilities) to meet its short-term debt obligations. In other words, it lets you know the resources available for a firm to use to pay its current obligation and expenses. If a company cannot maintain a short-term debt-paying ability, it will not be able to maintain a long-term debt-paying ability, nor will it be able to satisfy its stockholders. The main ratio in this group is the current ratio and acid test ratio.

The ***Current Ratio*** is a liquidity ratio that measures a company's ability to pay short-term and long-term obligations. To gauge this ability, the current ratio considers the current total assets of a company (both liquid and illiquid) relative to that company's current total liabilities.

Current Ratio is the most widely used of all analytical devices based on the balance sheet. It establishes the relationship between Current assets & Current liabilities.

Current Ratio = Current Assets / Current Liabilities

The current ratio is mainly used to give an idea of the company's ability to pay back its liabilities (debt and accounts payable) with its assets (cash, marketable securities, inventory, accounts receivable). As such, current ratio can be used to take a rough measurement of a company's financial health.

It compares a firm's current assets to its current liabilities, and is expressed as follows: The current ratio is an indication of a firm's liquidity. Acceptable current ratios vary from industry to industry.

Ideal Ratio: 2:1

High Ratio indicates under trading and over capitalization.

Low Ratio indicates over trading and under capitalization.

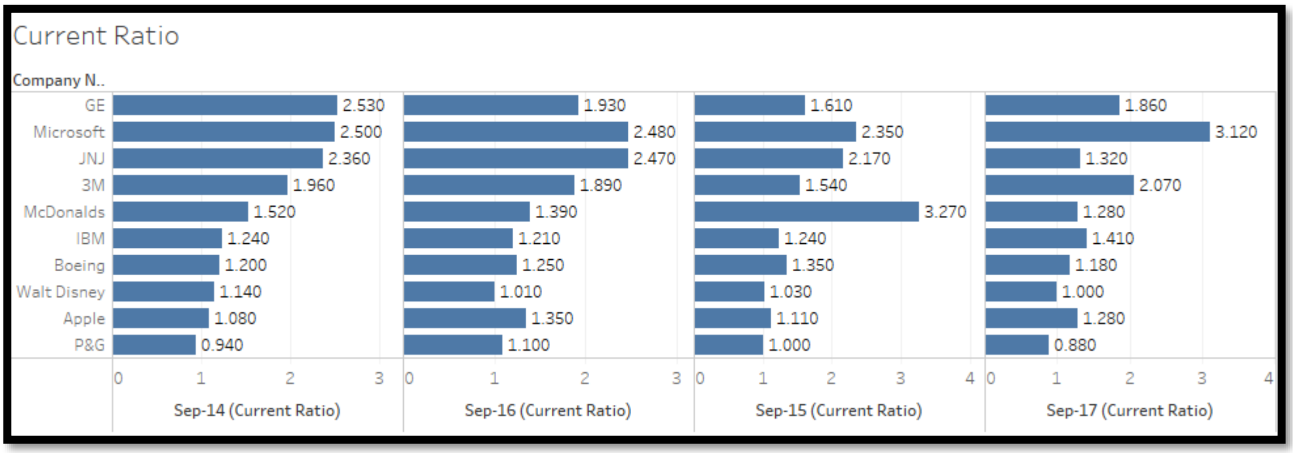


Fig 3: Current Ratio of different Stocks

<u>Leverage Ratio</u>:

A leverage ratio is any one of several financial measurements that look at how much capital comes in the form of debt (loans), or assesses the ability of a company to meet its financial obligations. The leverage ratio is important given that companies rely on a mixture of equity and debt to finance their operations, and knowing the amount of debt held by a company is useful in evaluating whether it can pay its debts off as they come due.

Too much debt can be dangerous for a company and its investors. However, if a company's operations can generate a higher rate of return than the interest rate on its loans, then the debt is helping to fuel growth in profits. Nonetheless, uncontrolled debt levels can lead to credit downgrades or worse. On the other hand, too few debts can also raise questions. A reluctance or inability to borrow may be a sign that operating margins are simply too tight.

The most well-known financial leverage ratio is the ***debt-to-equity ratio***.

D/E Ratio = Total Debt / Total Equity

The D/E ratio indicates how much debt a company is using to finance its assets relative to the value of shareholders' equity.

A high debt/equity ratio generally indicates that a company has been aggressive in financing its growth with debt. This can result in volatile earnings because of the additional interest expense. If the company's interest expense grows too high, it may increase the company's chances of a default or bankruptcy. Typically, a D/E ratio greater than 2.0 indicates a risky scenario for an investor however this yardstick can vary by industry.

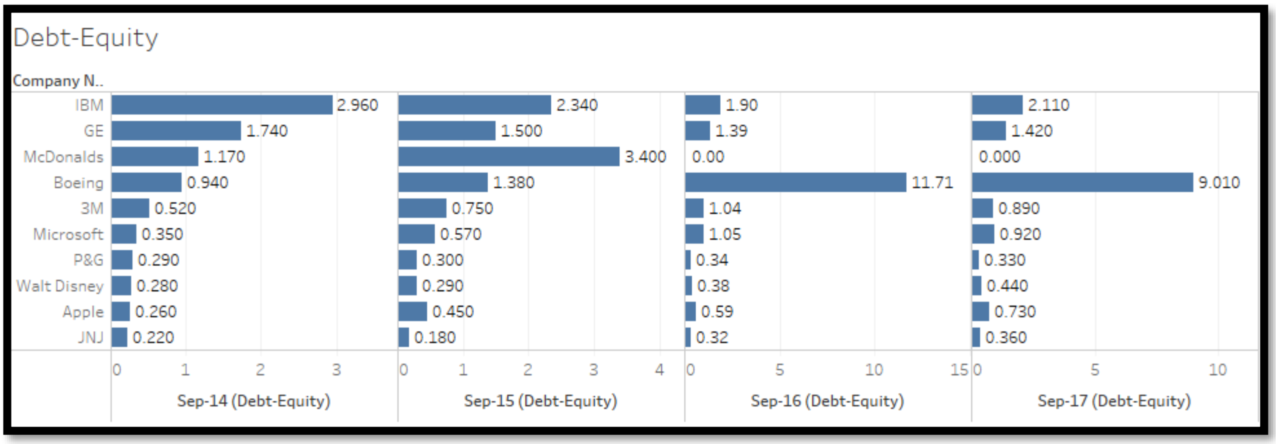


Fig 4: Debt-Equity Ratio of different Stocks

Profitability Ratio:

The profitability ratios group, also known as performance ratios, assesses the company ability to earn profits on sales, assets and equity, it measures the return earned on a company's capital and the financial cushion relative to each dollar of sales, These are critical to determining the attractiveness of investing in company shares, and investors in using these ratios widely, much like the operational performance ratios, these ratios give users a good understanding of how well the company utilized its resources(assets) in generating profit and shareholder value.

Some examples of profitability ratios are profit margin, return on assets (ROA) and return on equity (ROE). Profitability ratios are the most popular metrics used in financial analysis. The long-term profitability of a company is vital for both the survivability of the company as well as the benefit received by shareholders.

Return on Equity:

Return on Equity (ROE) is the amount of net income returned as a percentage of shareholder's equity. Return on equity measures a corporation's profitability by revealing how much profit a company generates with the money shareholders have invested.

ROE is expressed as a percentage and calculated as:

Return on Equity = Net Income/Shareholder's Equity

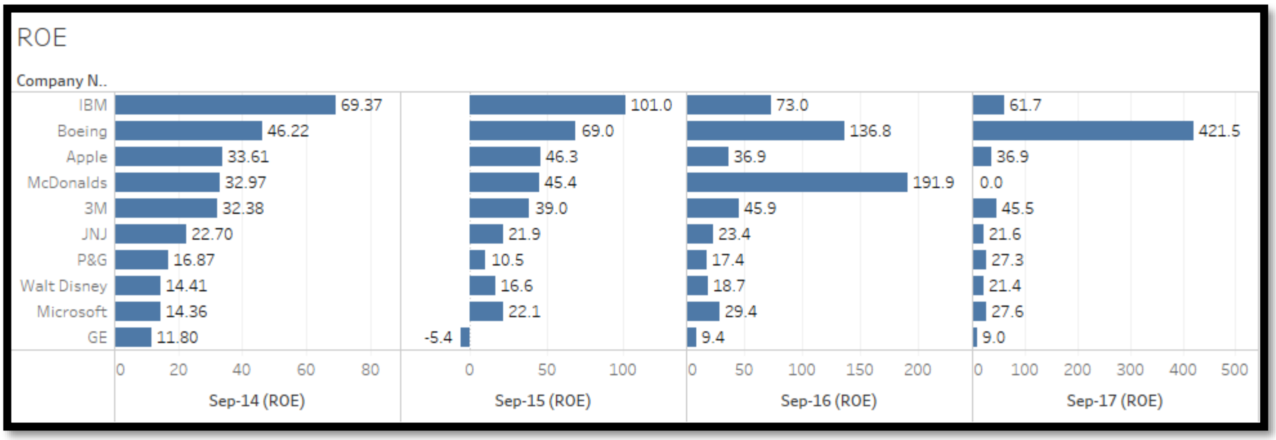


Fig 5: Return on Equity Ratio of different Stocks

Operating Profit Ratio:

This ratio establishes a relationship between cost of goods sold plus other operating expenses and net sales. This ratio is calculated mainly to ascertain the operational efficiency of the management in their business operations.

Operating ratio = (Cost of goods sold + operating expenses) / Net sales

Higher the ratio the less favorable it is because it would leave a smaller margin to meet interest, dividend and other corporate needs. This ratio is partial index of overall profitability.

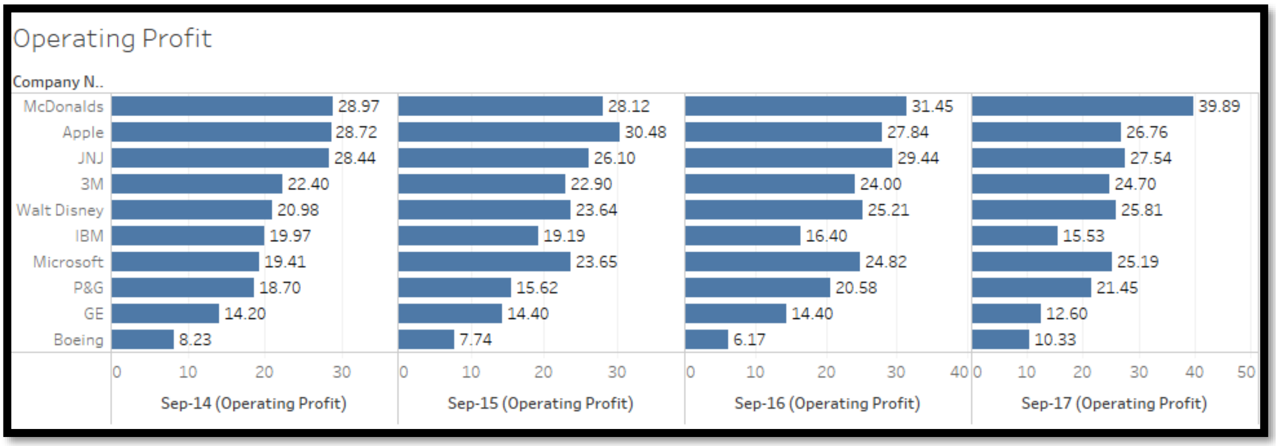


Fig 6: Operating Profit Ratio of different Stocks

Activity Ratio:

Activity ratios measure a firm's ability to convert different accounts within its balance sheet into cash or sales. Activity ratio measure the relative efficiency of a firm based on its use of its assets, leverage or other such balance sheet items and are important in determining whether a company's management is doing a good enough job of generating revenues and cash from its resources.

Activity ratios gauge an organization's operational efficiency and profitability. Activity ratios are most useful when compared to competitor or industry to establish whether an entity's processes are favorable or unfavorable. Activity ratios can form a basis of comparison across multiple reporting periods to determine changes over time.

Asset Turnover Ratio:

Assets turnover ratio take a look at how efficiently an entity uses its assets to make a sale. Total sales are divided by total assets to see how proficient a business is at using its assets. Smaller ratios may indicate that the company is holding higher levels of inventory instead of selling.

Asset Turnover = Sales / Average Total Assets

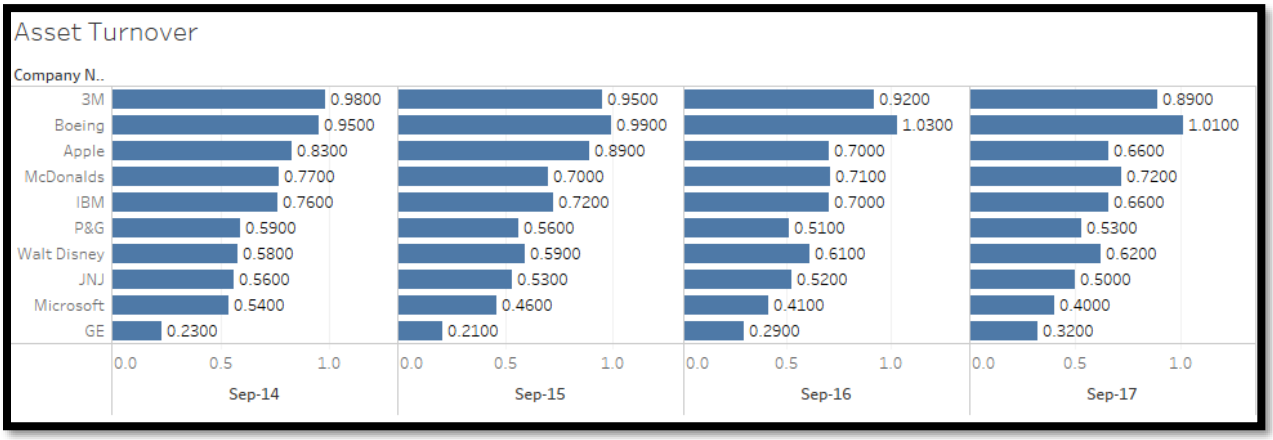


Fig 7: Asset Turnover Ratio of different Stocks

Chapter 3

Data under Study

Financial Data

The volume and variety of data used by stock market investors have exploded in recent years. Gone are the days when stock prices and company fundamentals sufficed to build a robust investment strategy.

Today, analysts and traders use a far wider set of data to inform their investment decisions: market structure data, sentiment and psychology data, consensus and survey data and much more.

Quandl has the richest collection of publicly available equity data on the internet. Quandl offers a simple API for stock market data downloads. Our daily data feeds deliver end-of-day prices, historical stock fundamental data, harmonized fundamentals, financial ratios, indexes, options and volatility, earnings estimates, analyst ratings, investor sentiment and more.

Quandl's data products come in many forms and contain various objects, including time-series and tables. Through APIs and various tools (R, Python, Excel, etc.), users can access/call the premium data to which they have subscribed.

We are studying the fundamental stock ratios available in Quandl using Python API.

Reliable stock quotes are the foundation of any trading strategy. Analysts require accurate current stock quotes, to judge prevailing market conditions and bias-free historical stock prices, for back testing and research.

Stock Quotes

There are a multitude of stock quote providers; however, not all of them are equivalent in quality. At a minimum, a good stock data publisher should have excellent, robust, transparent methodology for data collection, ideally directly from primary sources (exchanges). Historical stock prices need to be correctly adjusted for splits, dividends, spinoffs, mergers and other corporate actions.

Comprehensive documentation is a must: You should ask how the data is sourced, how it is organized, how it is adjusted if necessary, and how it can be accessed by the user. The data itself must be pristine: free of errors, gaps and outliers. Timeliness, reliability and consistency are essential for professional applications. Responsive customer support is another way that providers can add value to their data.

Surprisingly few data publishers satisfy all these criteria; nonetheless, the best stock market analysts insist on these. Quandl has aggregated several databases from a variety of publishers that score highly on all these components. These publishers cover both current and historical stock quotes, spanning different exchanges and stock classes. A list of different types of stock quotes follows:

The following five end-of-day stock price databases cover a single exchange each, with comprehensive quotes for that exchange.

- NYSE Stock Prices – End-of-day and historical quotes for all 11,500 stocks traded on the New York Stock Exchange.
- NASDAQ Stock Prices – End-of-day and historical prices for all 28,500 NASDAQ stocks.
- NYSE MKT (AMEX) Prices – Historical and end-of-day quotes for all 12,000 NYSE MKT stocks (formerly American Stock Exchange).

- OTC Bulletin Board Prices – Prices for 12,000 over-the-counter bulletin board stocks, with history going back to 2007.
- OTC Markets Group Prices – Current and historical quotes for over 48,000 OTC Markets Group stocks (formerly Pink Sheets).

The above five stock market databases all include daily quotes, historical prices, dividends, splits and other adjustments.

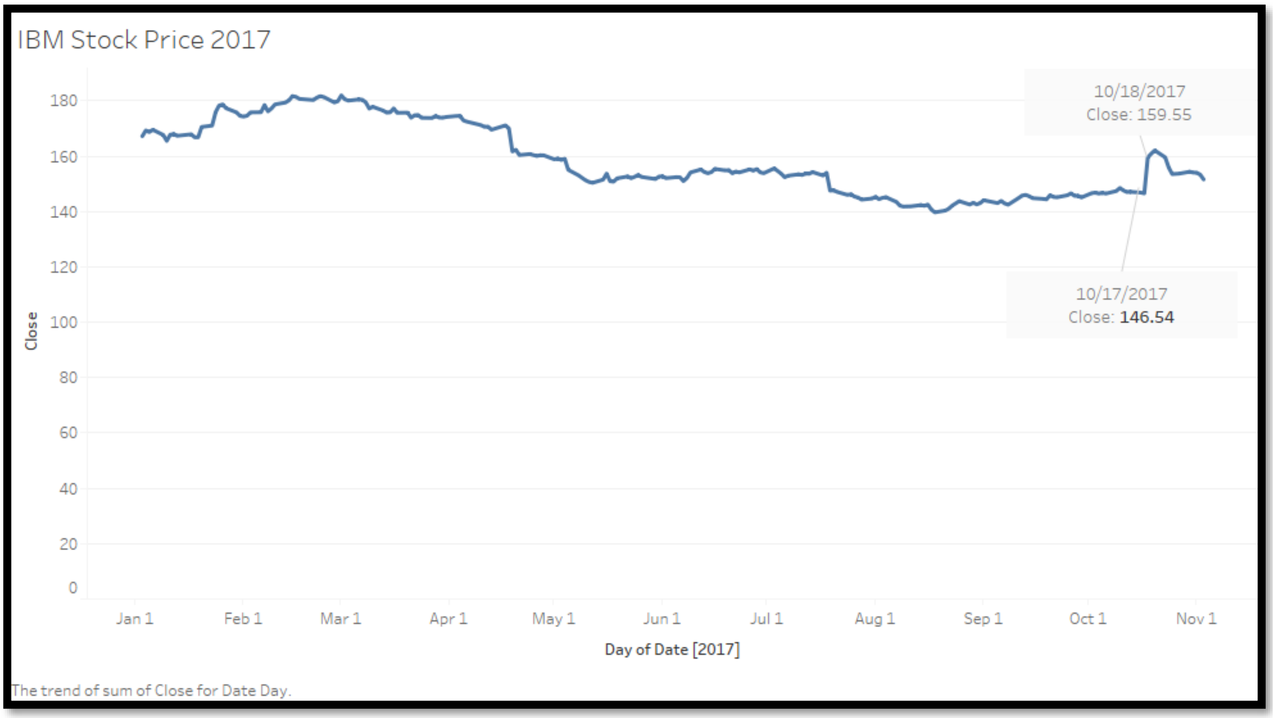


Fig 8: IBM Stock Price change in 2017

Sentiment Data

For performing sentimental analysis, we have collected data from Twitter related to the stocks that we have selected.

Tweets are classified as positive, negative and neutral based on the sentiment present. Approx. 1000 tweets out of the total tweets are examined manually and scaled as 1 for Extremely negative, 2 for Negative, 3 for Neutral, 4 for Positive and 5 for Extremely positive emotions. For classification of other tweets a machine learning model is trained whose features are extracted from the manually scaled tweets.

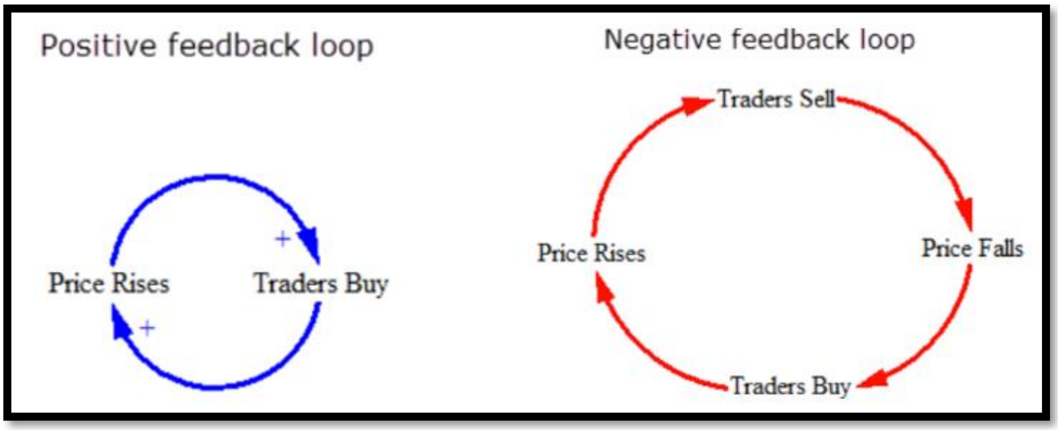


Fig 9: Postive – Negative sentiments

Analytical Approach

The stock market is tricky. It's one of the more difficult predictive challenges; there are inherent numerical patterns as well as a wide range of interactions with the outside world. As a market, it is continuously adjusting to the latest prediction theories in continuous coevolution.

With the advent of the digital computer, stock market prediction has since moved into the technological realm. The most prominent technique involves the use of artificial neural networks (ANNs) and Genetic Algorithms. The most common form of NN in use for stock market prediction is the feed forward network utilizing the backward propagation of errors algorithm to update the network weights. These networks are commonly referred to as Backpropagation networks. Another form of ANN that is more appropriate for stock prediction is the time recurrent neural network (RNN) or time delay neural network (TDNN).

There are usually two approaches taken for forecasting different time horizons: independent and joint. The independent approach employs a single NN for each time horizon, for example, 1-day, 2-day, or 5-day. The advantage of this approach is that network forecasting error for one horizon won't impact the error for another horizon—since each time horizon is typically a unique problem. The joint approach, however, incorporates multiple time horizons together so that they are determined simultaneously. In this approach, forecasting error for one time horizon may share its error with that of another horizon, which can decrease performance. There are also more parameters required for a joint model, which increases the risk of overfitting.

The use of Text Mining together with Machine Learning algorithms received more attention in the last years, with the use of textual content from Internet as input to predict price changes in Stocks and other financial markets. The collective mood of Twitter messages has been linked to stock market performance.

The model that we have right now consist of several stages:

- First, we perform adaptive filtering of the raw price data
- Then we feed filtered data into the NN model and train it

- Then we perform predictions and prepare the recommendations using those predictions.

Few highlights:

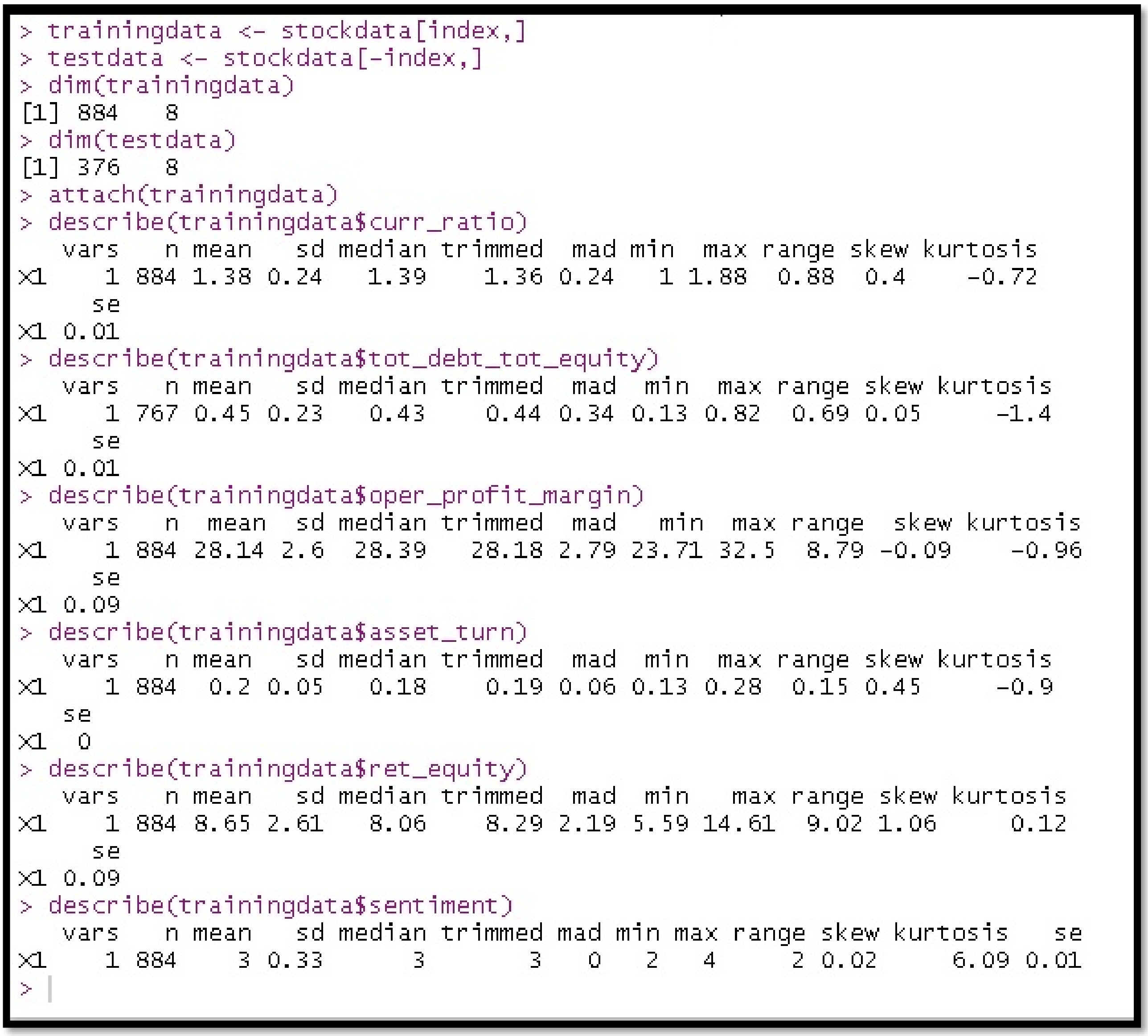
- A data analytic approach is proposed for stock market prediction.
- The proposed decision support system effectively uses fewer variables.
- Higher prediction accuracy is achieved compared to the literature.
- Sophisticated data analytics methods present capability to capture nonlinearity.
- This generic methodology is applicable to the other emerging markets as well.

Exploratory Data Analysis

In case of Financial data, we did not face any discrepancies as the data was clean.

Tweets consists of more noisy words such as many numbers, more white spaces, tabs, punctuation characters, stop words, acronyms, emoticons and unnecessary data like pictures and URL's. So, tweets are pre-processed to represent correct emotions of public. For pre-processing of tweets, we employed three stages of filtering: Tokenization, stop words removal and regex matching for removing special characters.

We have picked few sample stocks for EDA and here are the findings.



```
> trainingdata <- stockdata[index,]
> testdata <- stockdata[-index,]
> dim(trainingdata)
[1] 884   8
> dim(testdata)
[1] 376   8
> attach(trainingdata)
> describe(trainingdata$curr_ratio)
   vars   n mean   sd median trimmed  mad min  max range skew kurtosis
X1    1 884 1.38 0.24   1.39    1.36 0.24   1 1.88  0.88  0.4    -0.72
     se
X1 0.01
> describe(trainingdata$tot_debt_tot_equity)
   vars   n mean   sd median trimmed  mad  min  max range skew kurtosis
X1    1 767 0.45 0.23   0.43    0.44 0.34 0.13 0.82  0.69 0.05     -1.4
     se
X1 0.01
> describe(trainingdata$oper_profit_margin)
   vars   n  mean  sd median trimmed  mad   min  max range  skew kurtosis
X1    1 884 28.14 2.6  28.39   28.18 2.79 23.71 32.5  8.79 -0.09    -0.96
     se
X1 0.09
> describe(trainingdata$asset_turn)
   vars   n mean   sd median trimmed  mad  min  max range skew kurtosis
X1    1 884  0.2 0.05   0.18    0.19 0.06 0.13 0.28  0.15 0.45     -0.9
   se
X1  0
> describe(trainingdata$ret_equity)
   vars   n mean   sd median trimmed  mad  min   max range skew kurtosis
X1    1 884 8.65 2.61   8.06    8.29 2.19 5.59 14.61  9.02 1.06     0.12
     se
X1 0.09
> describe(trainingdata$sentiment)
   vars   n mean   sd median trimmed mad min max range skew kurtosis   se
X1    1 884    3 0.33      3       3   0   2   4     2 0.02     6.09 0.01
> 
```

Fig 10: Exploratory Data Analysis

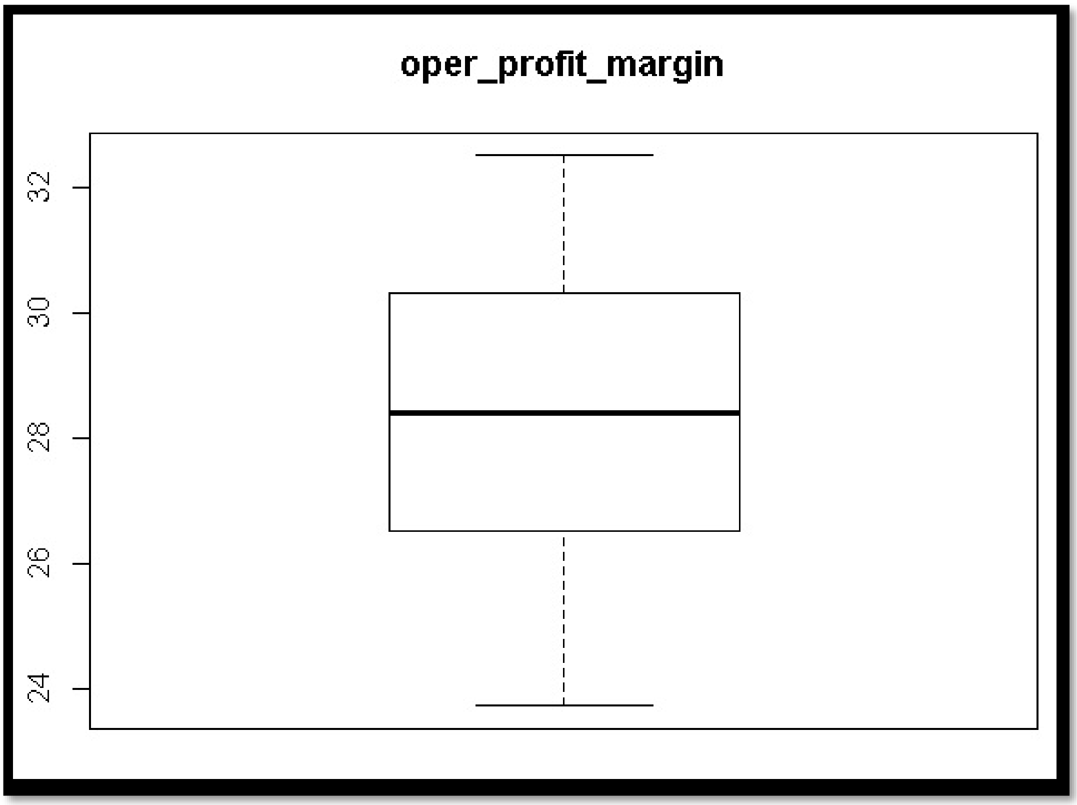


Fig 11: Operational Profit Margin

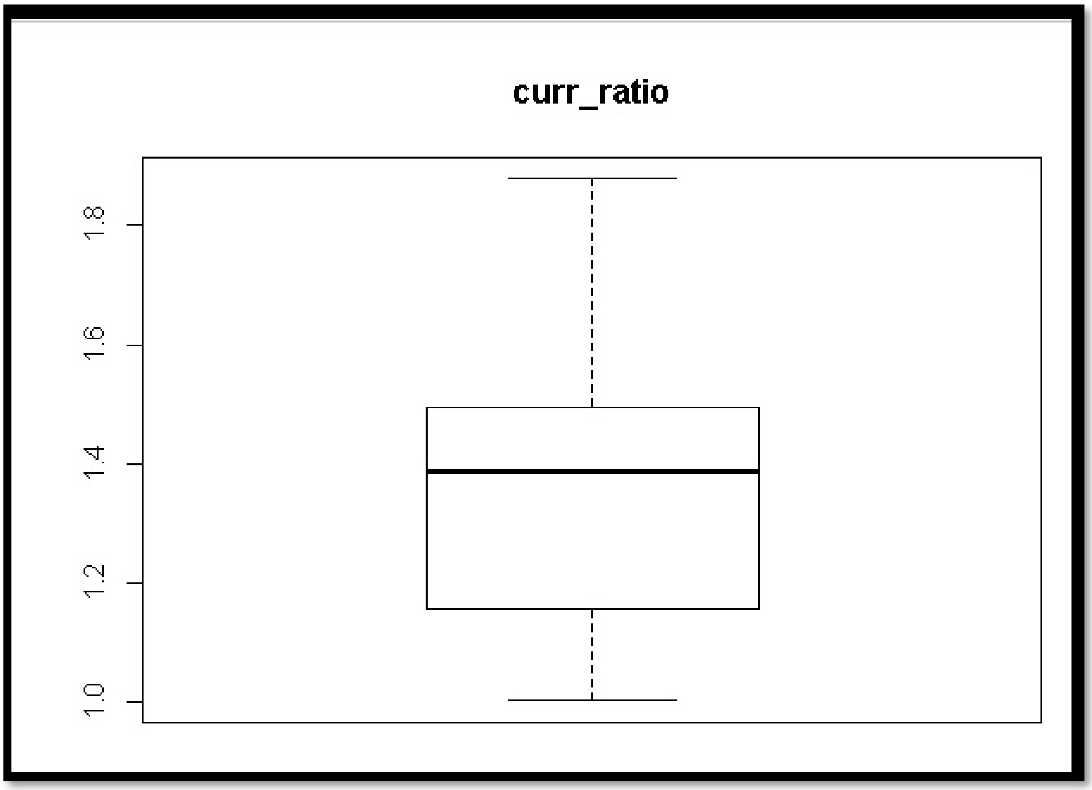


Fig 12: Current Ratio

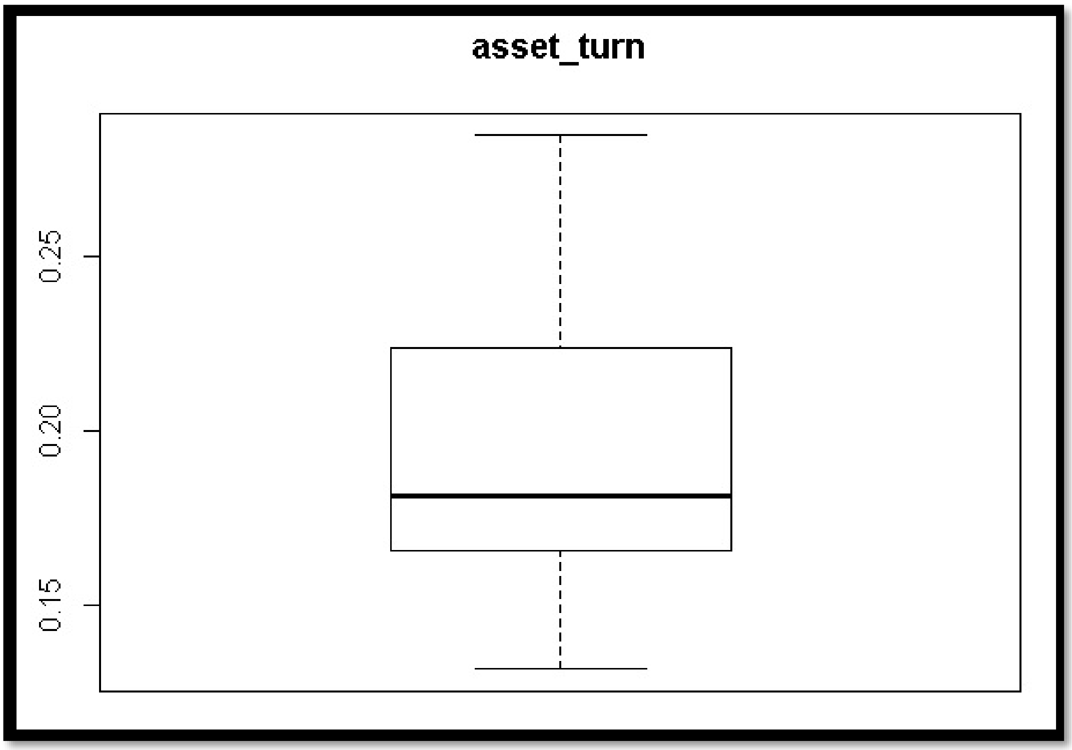


Fig 13: Asset Turn

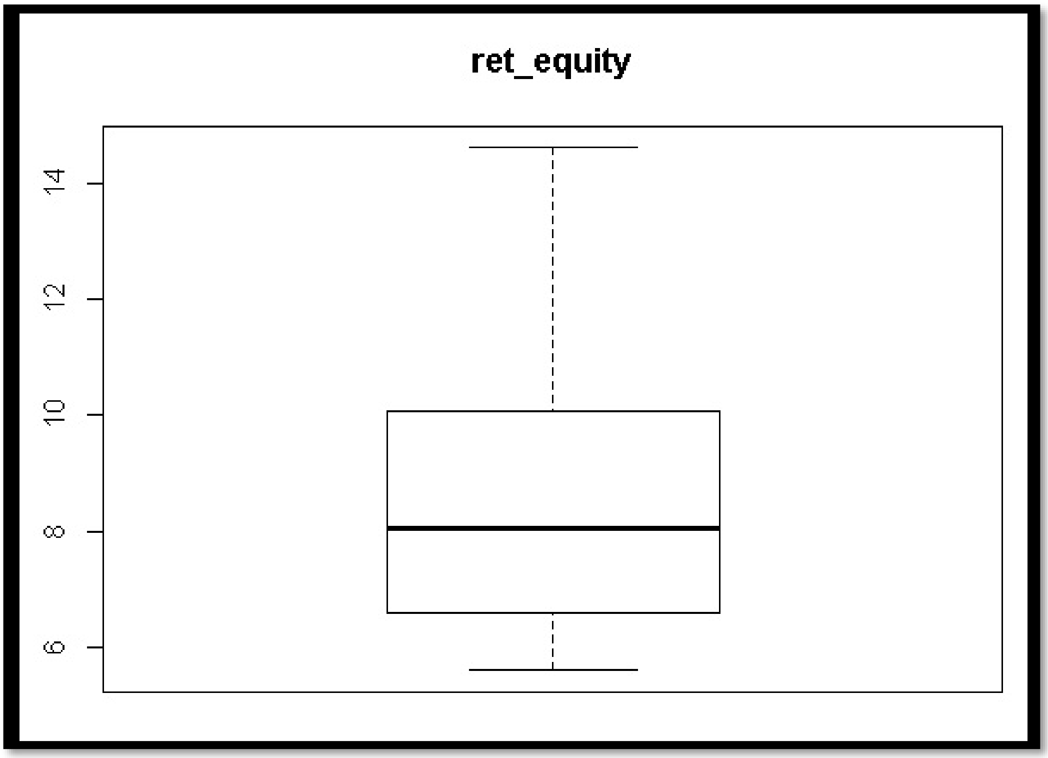


Fig 14: Return on Equity

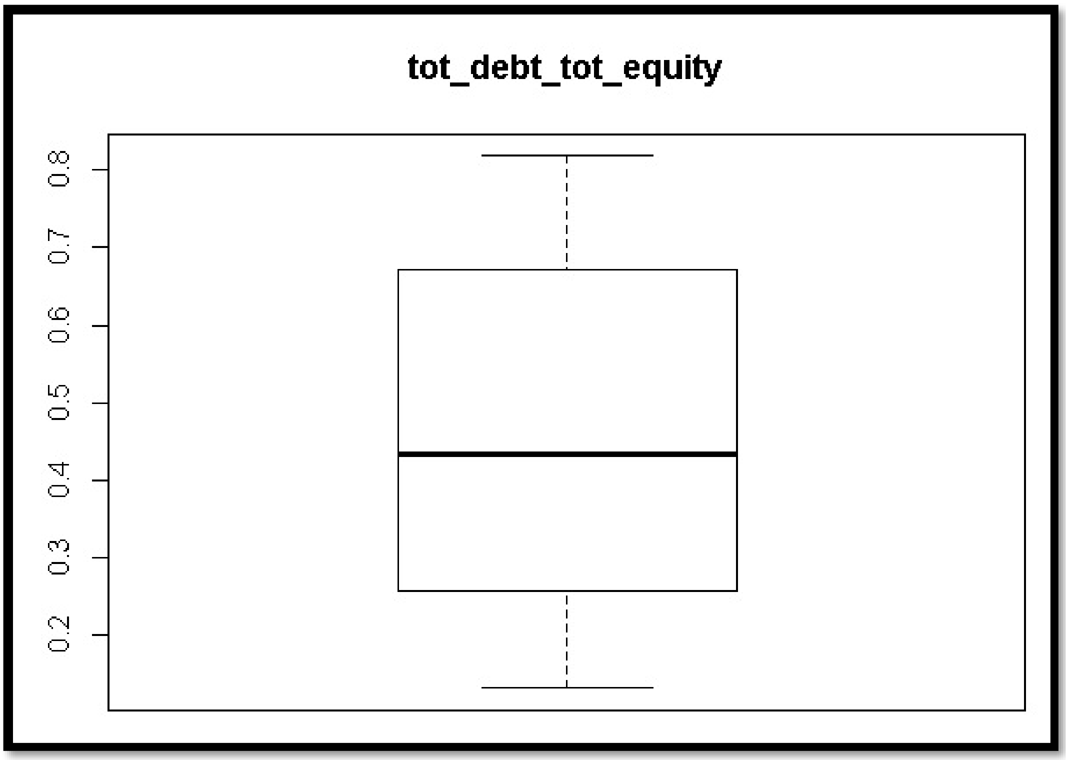


Fig 15: Total Debt to Total Equity

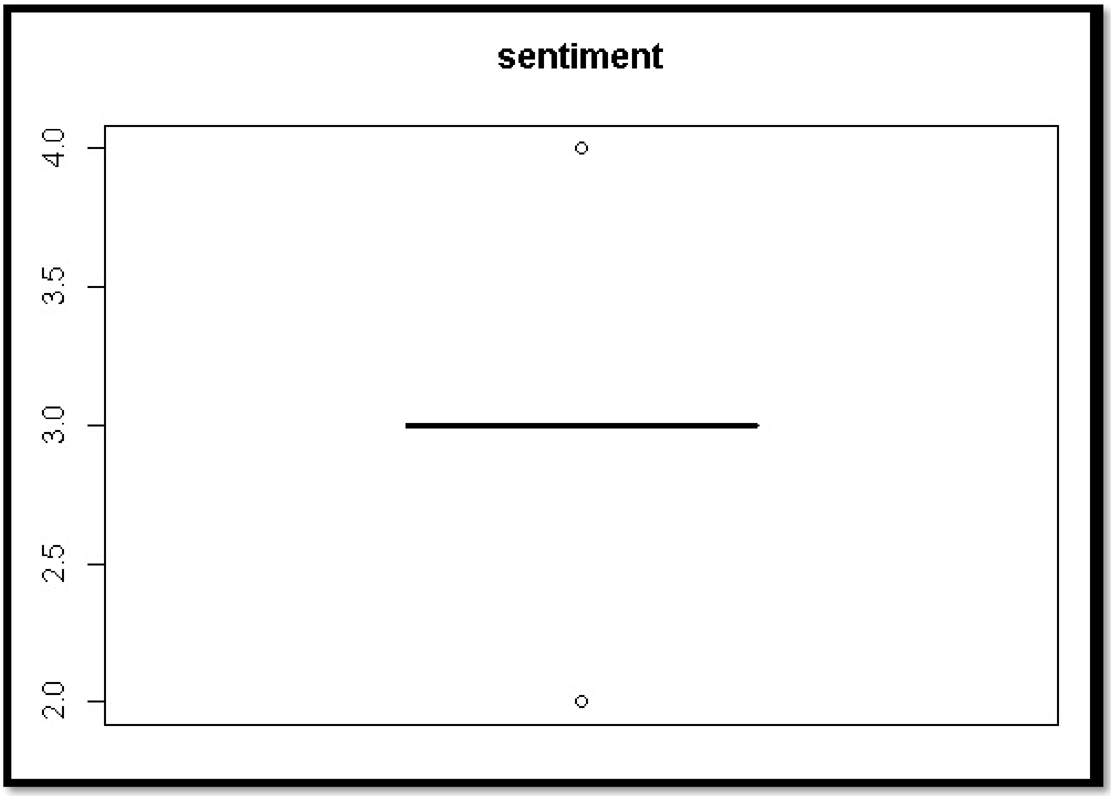


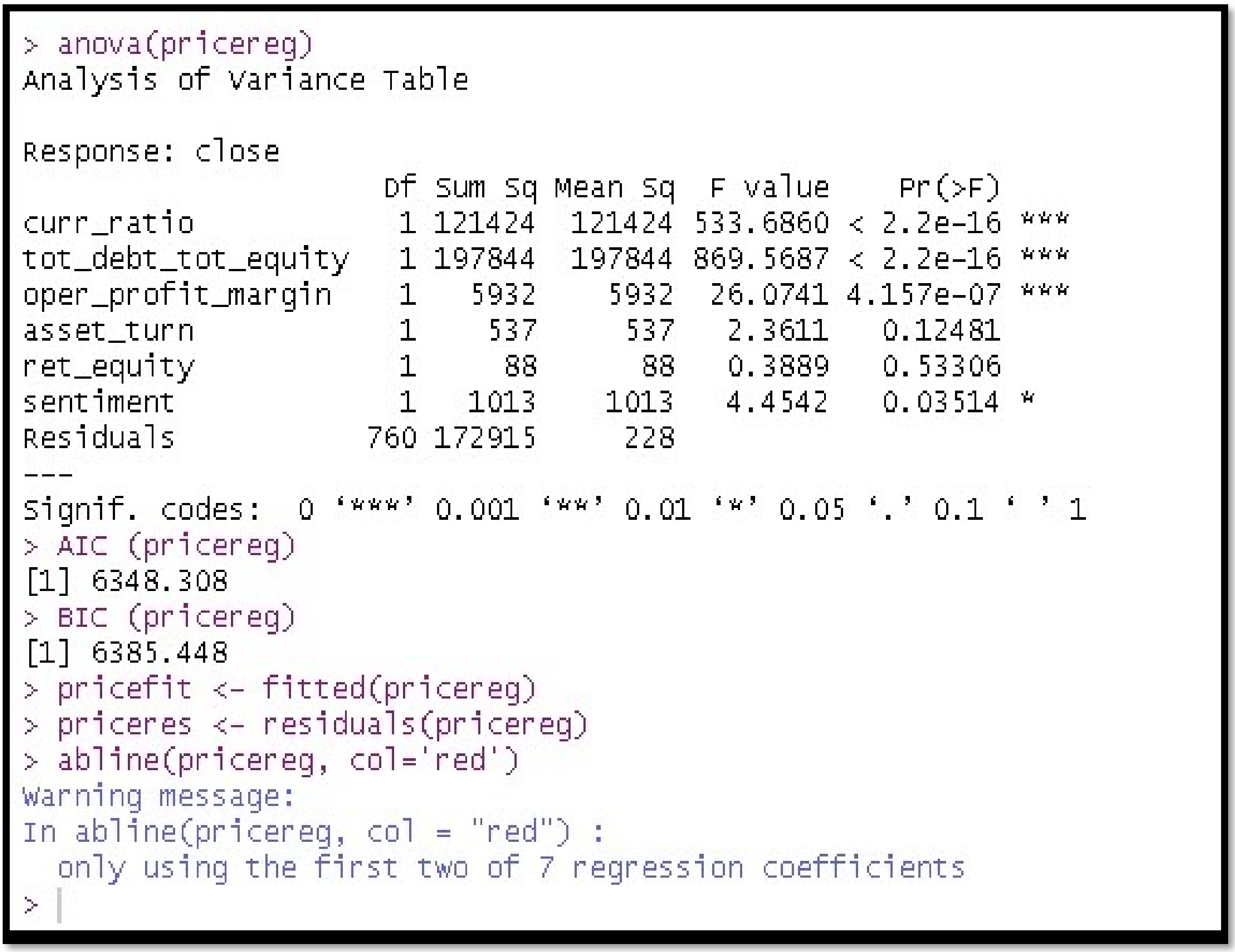
Fig 16: Sentiment

Regression Model:

We define Null and Alternative hypothesis as below and prove that Alternative Hypothesis holds good.

Null hypothesis, H_0: all predictor variables like ratios, sentiments have NO impact on the price change.

Alternate hypothesis, H_a: all predictor variables like ratios, sentiments have impact on the price change.



```
> anova(pricereg)
Analysis of Variance Table

Response: close
                     Df Sum Sq Mean Sq  F value    Pr(>F)    
curr_ratio            1 121424  121424 533.6860 < 2.2e-16 ***
tot_debt_tot_equity   1 197844  197844 869.5687 < 2.2e-16 ***
oper_profit_margin    1   5932    5932  26.0741 4.157e-07 ***
asset_turn            1    537     537   2.3611   0.12481    
ret_equity            1     88      88   0.3889   0.53306    
sentiment             1   1013    1013   4.4542   0.03514 *  
Residuals           760 172915     228                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> AIC (pricereg)
[1] 6348.308
> BIC (pricereg)
[1] 6385.448
> pricefit <- fitted(pricereg)
> priceres <- residuals(pricereg)
> abline(pricereg, col='red')
Warning message:
In abline(pricereg, col = "red") :
  only using the first two of 7 regression coefficients
> |
```

Fig 17: Results from regression model

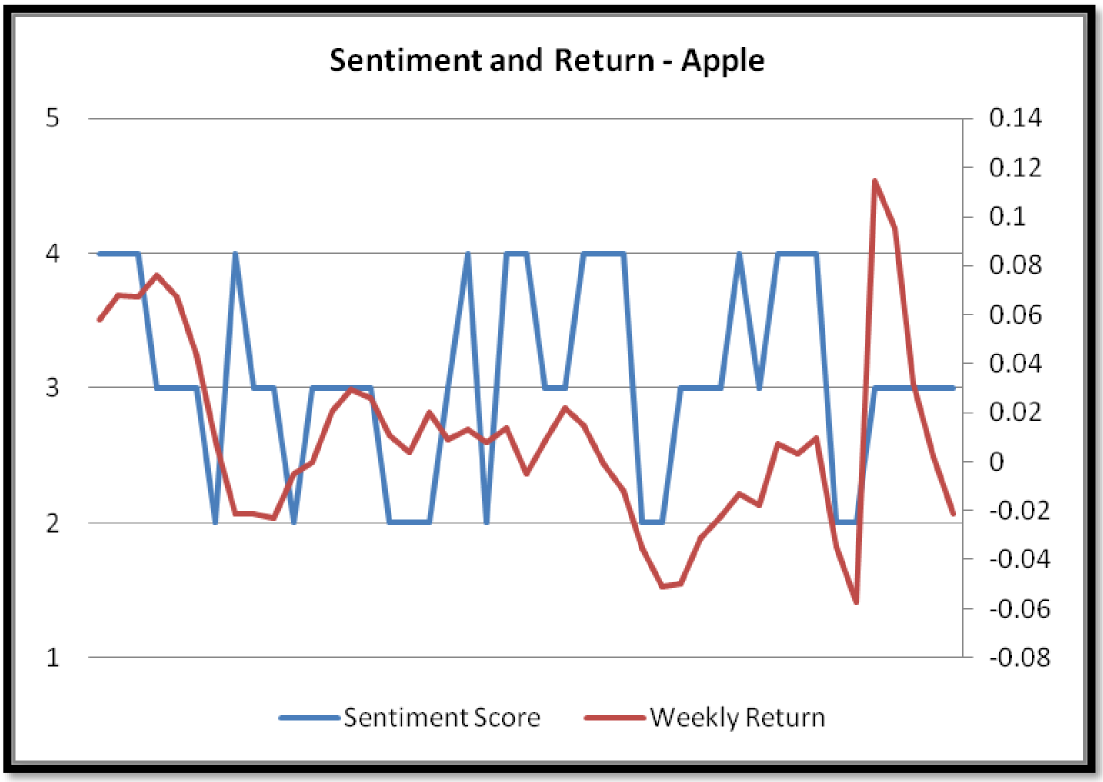


Fig 18: Sentiment Score with return Apple

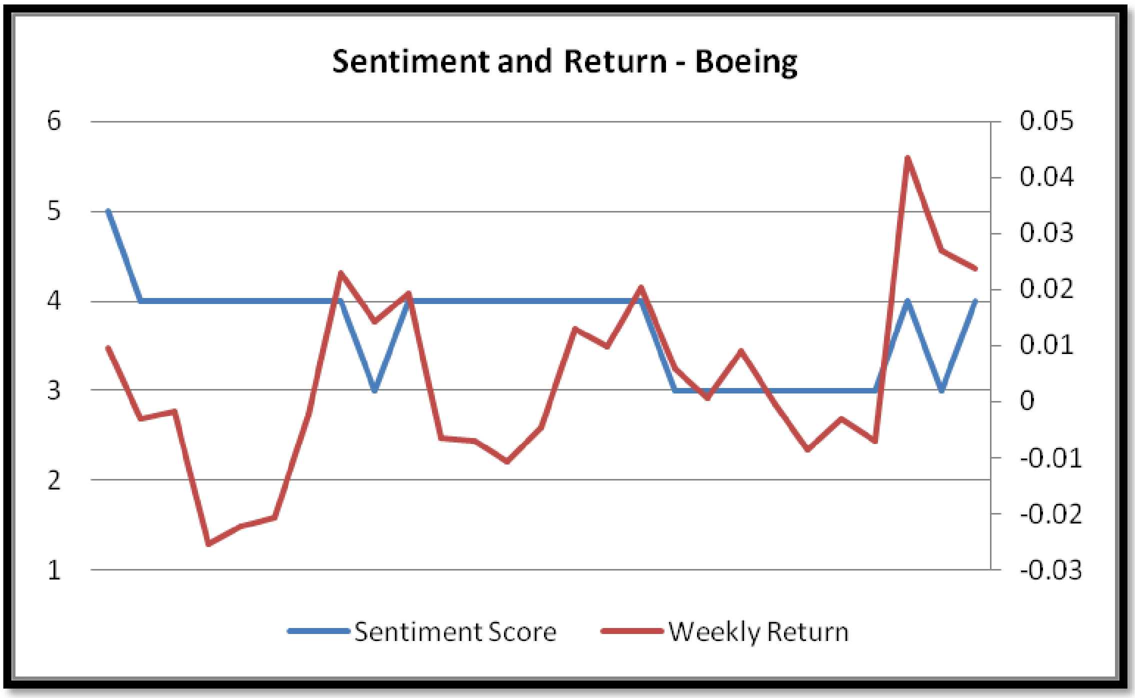


Fig 19: Sentiment Score with return Boeing

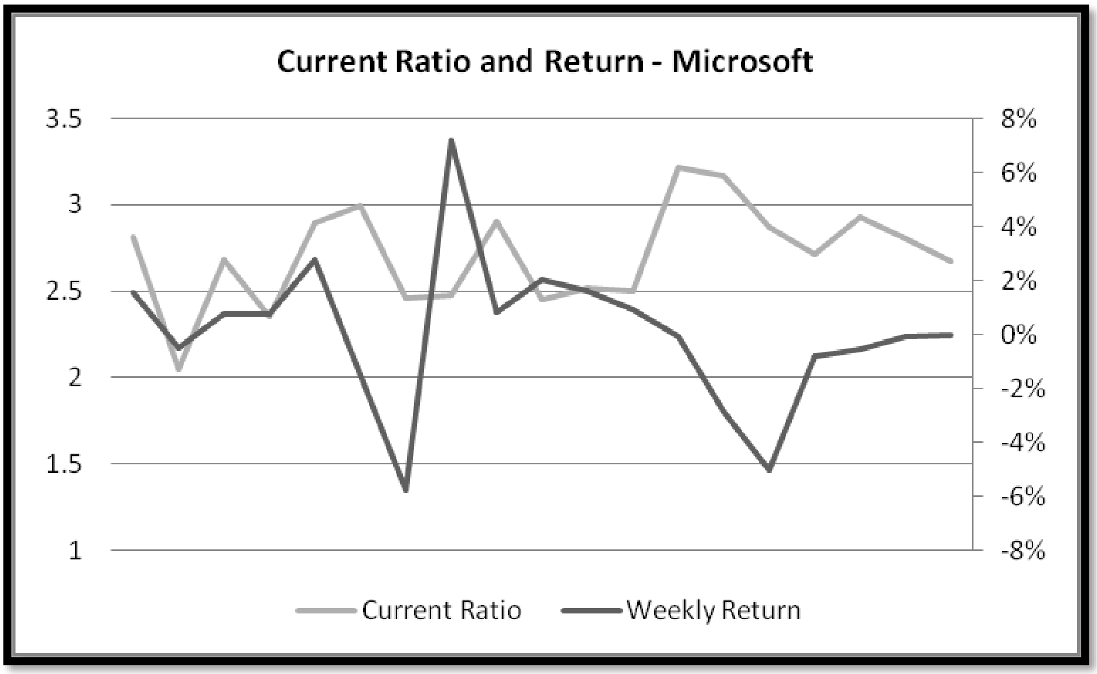


Fig 20: Current Ratio with return

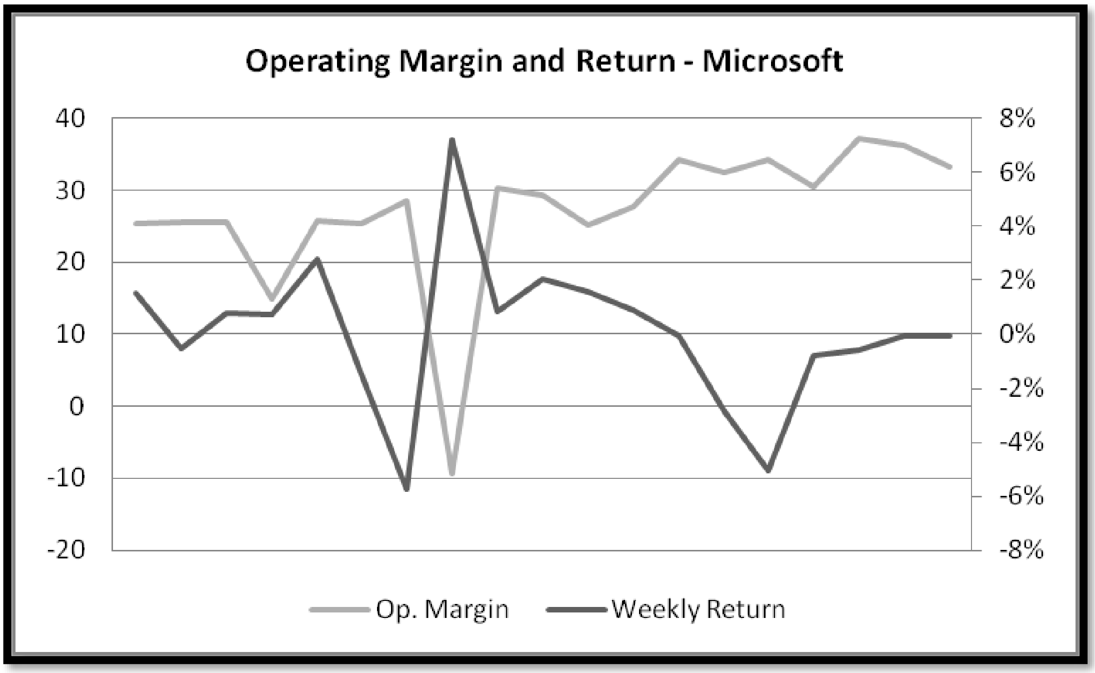


Fig 21: Operating Margin with return

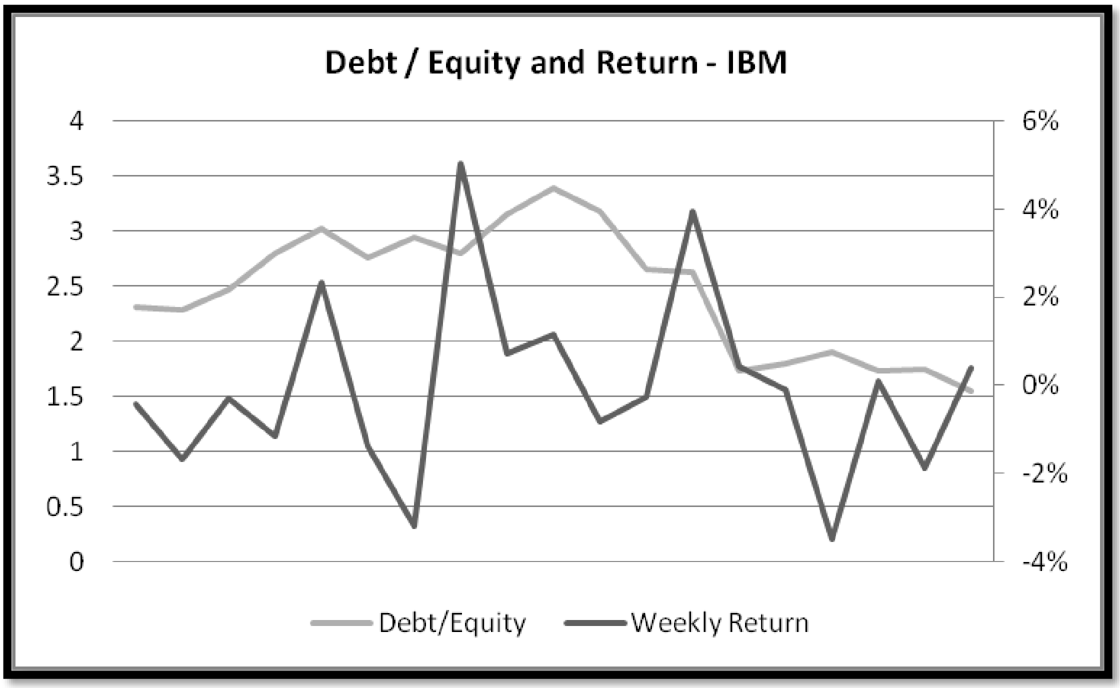


Fig 22: Debt Equity Ratio with return

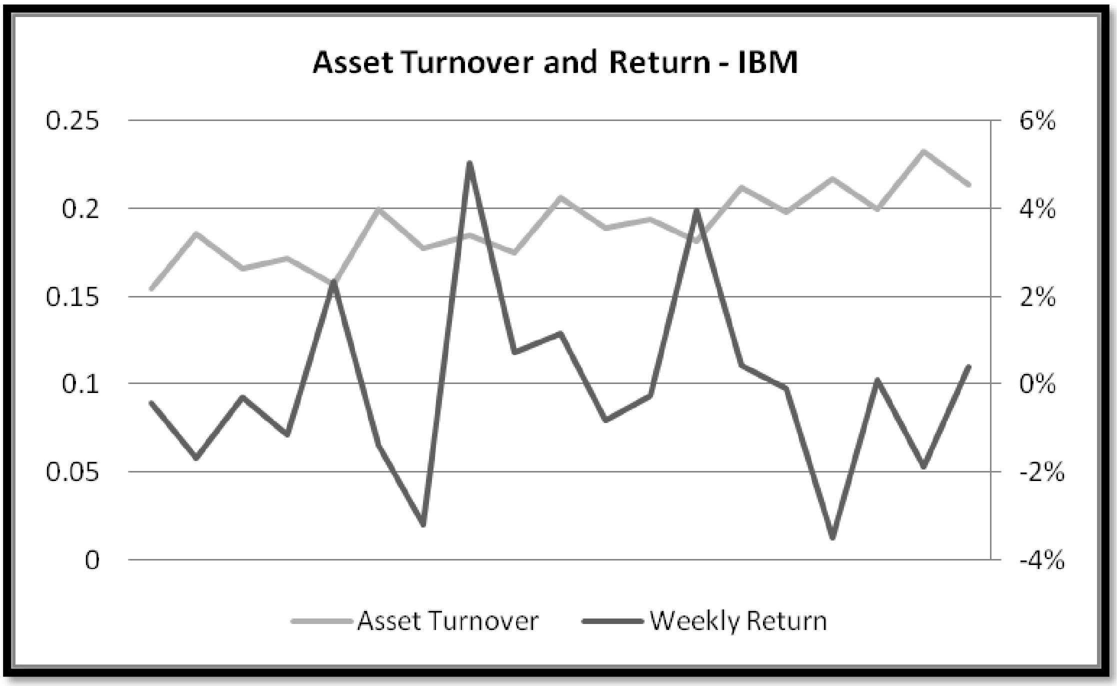


Fig 23: Asset Turnover with return

Chapter 4

Sentiment Analysis

Sentiment analysis is tricky and very much domain specific. There is lot of research on sentiment analysis of company performance and news articles and many sentiment analysers are available as an open source. The main problem with these analysers is that they are trained with a different corpus. For instance, Movie corpus and stock corpus are not equivalent. So, we developed our own sentiment analyser.

For performing sentimental analysis, we have collected data from Twitter related to the stocks that we have selected.

Tweets are classified as positive, negative and neutral based on the sentiment present. Approx. 1000 tweets out of the total tweets are examined manually and scaled as 1 for Extremely negative, 2 for Negative, 3 for Neutral, 4 for Positive and 5 for Extremely positive emotions. For classification of other tweets a machine learning model is trained whose features are extracted from the manually scaled tweets.

Sl.No	Tweet	Sentiment Score
1	b".@Apple co-founder's Woz U aims to train tech workers	4
2	b"'Anemic' iPhone 8 demand drags Apple shares lower	1
3	b"Apple CEO Tim Cook says rumored AR glasses won't be ready 'any time soon'	2
4	b"Apple CEO Tim Cook talks about new Chicago flagship store: 'A place where people can connect'	4
5	b"Apple CEO Tim Cook: Learn to code, it's more important than English as a second language	3
6	b"Apple co-founder Steve Wozniak: iPhone X is the first iPhone I won't buy on 'day one'	2
7	b"Apple co-founder Steve Wozniak: iPhone X is the first iPhone I won't buy on 'day one' 2	
8	b"Apple co-founder Steve Wozniak: There\xe2\x80\x99s \xe2\x80\x98way too much hype' around Elon Musk	3
9	b"Apple co-founder's Woz U aims to train tech workers	4
10	b"Apple COO Jeff Williams on tech and health care: 'we can't think of anything more significant' 4	
11	b"Apple COO Jeff Williams on tech and health care: 'we can't think of anything more significant'	4
12	b"Apple design chief Jony Ive says he despises 'most things'	4

13	b"Apple has agreed to terms with Bain Capital in the $18 billion bid for Toshiba's memory chip unit	4
14	b"Apple hit with trademark lawsuit over iPhone X 'animoji' feature	2

Table 1: Sample tweets downloaded for Apple

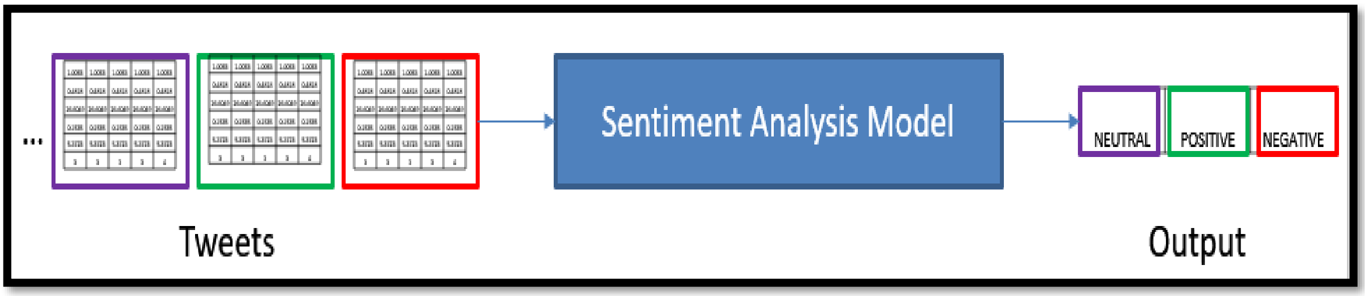
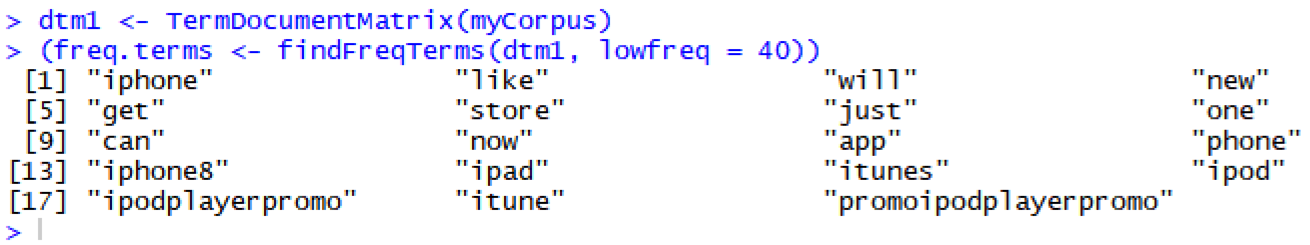


Fig 24: Sentiment Analysis Model flow

<u>Sentimental analysis R code and results:</u>



```
> dtm1 <- TermDocumentMatrix(myCorpus)
> (freq.terms <- findFreqTerms(dtm1, lowfreq = 40))
 [1] "iphone"           "like"             "will"             "new"
 [5] "get"              "store"            "just"             "one"
 [9] "can"              "now"              "app"              "phone"
[13] "iphone8"          "ipad"             "itunes"           "ipod"
[17] "ipodplayerpromo"  "itune"            "promoipodplayerpromo"
> |
```

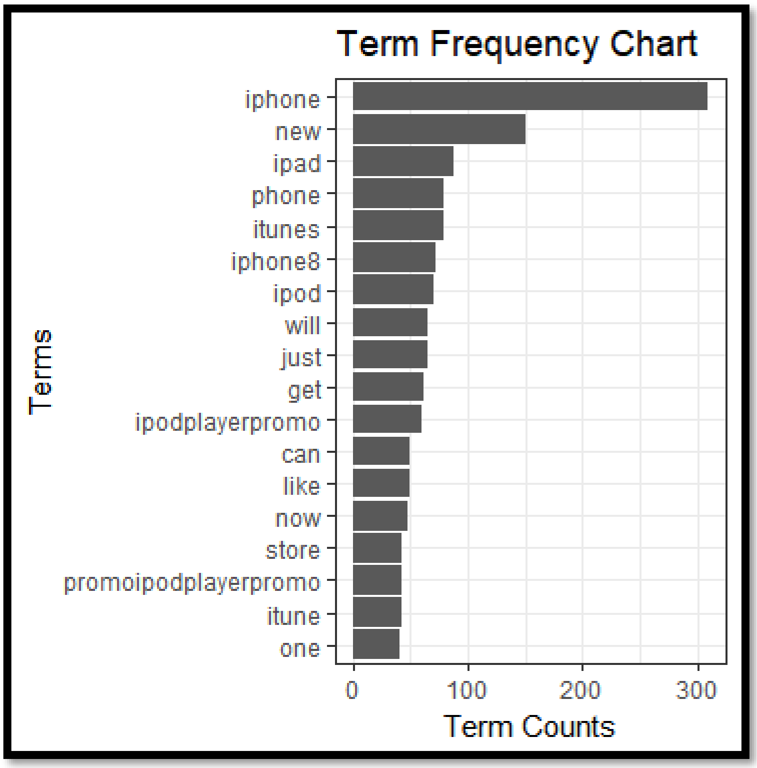


Fig 25: Term Frequency Chart

Word Cloud:

wordcloud(d$word,d$freq,min.freq=85)

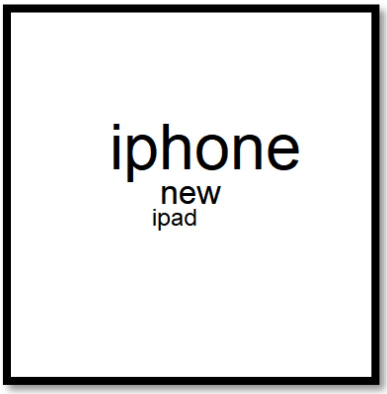


Fig 26: Word Cloud frequency = 85

wordcloud(d$word,d$freq,min.freq=65)

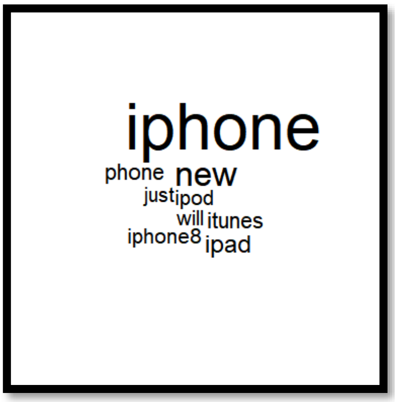


Fig 27: Word Cloud frequency = 65

wordcloud(d$word,d$freq,min.freq=45)

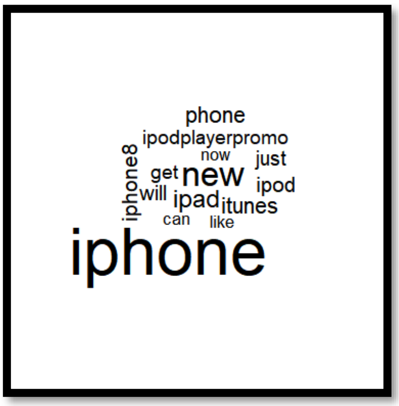
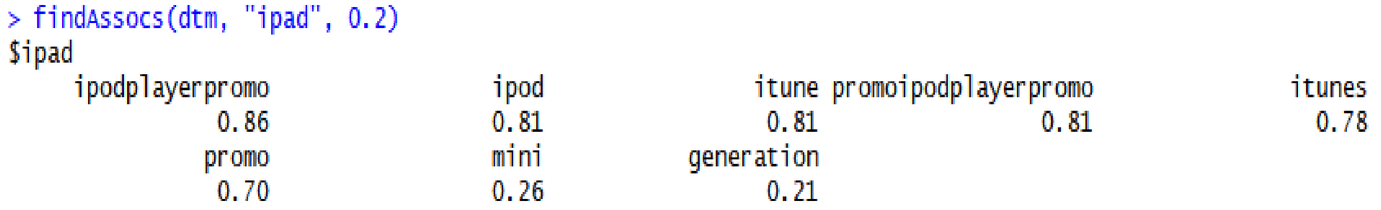
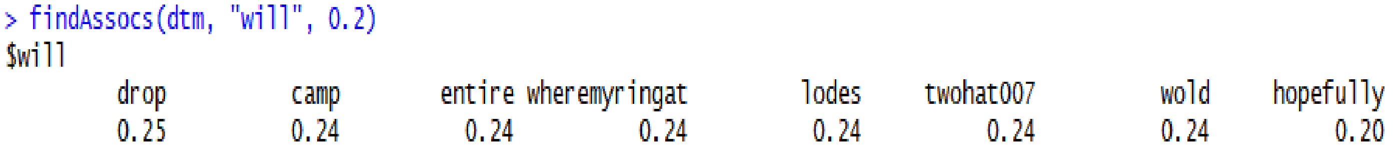


Fig 28: Word Cloud frequency = 45

Associations:

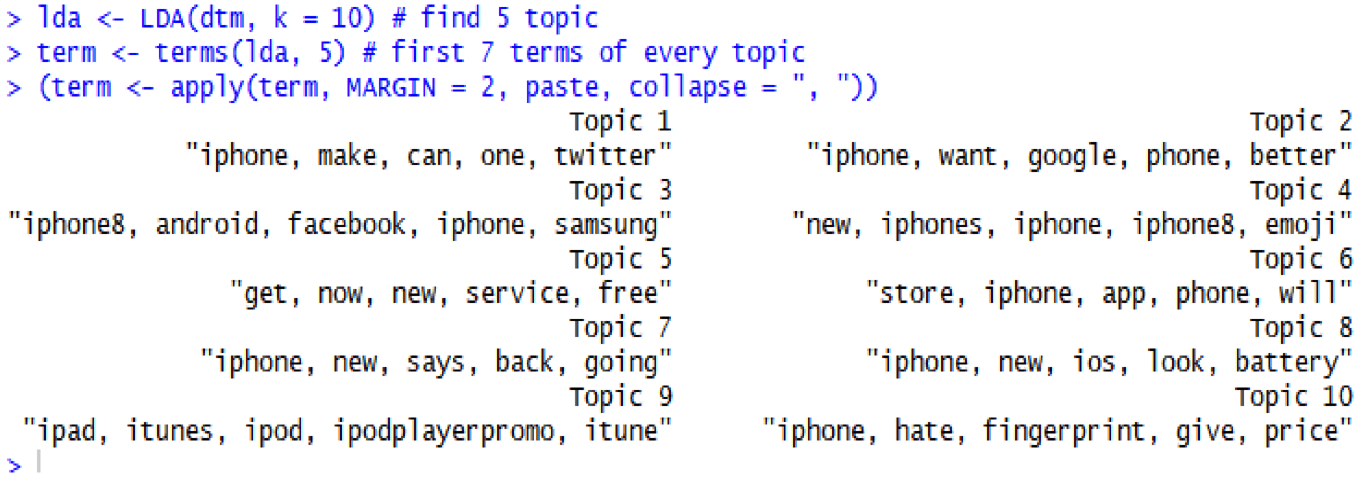


```
> findAssocs(dtm, "ipad", 0.2)
$ipad
     ipodplayerpromo              ipod             itune promoipodplayerpromo              itunes
                0.86              0.81              0.81                 0.81                0.78
               promo              mini        generation
                0.70              0.26              0.21
```

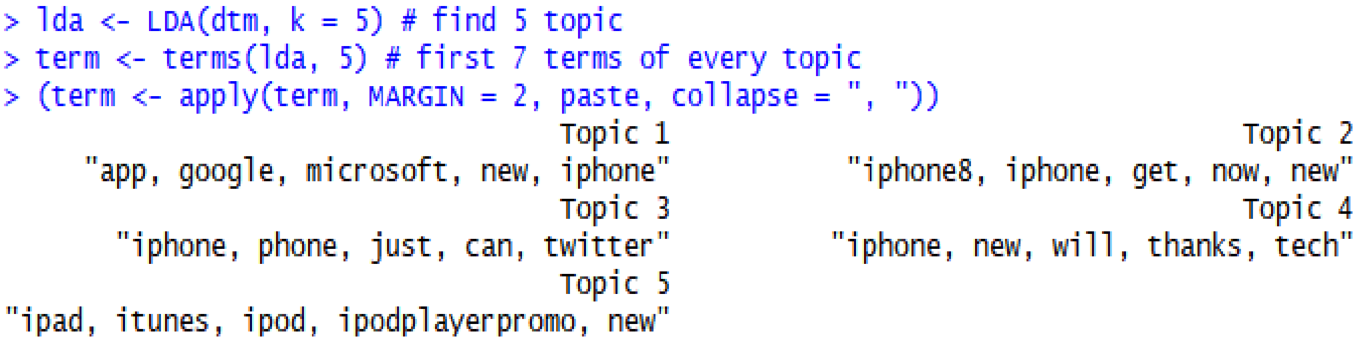


```
> findAssocs(dtm, "will", 0.2)
$will
        drop         camp       entire wheremyringat        lodes     twohat007         wold     hopefully
        0.25         0.24         0.24          0.24         0.24          0.24         0.24          0.20
```

Topic modelling:



```
> lda <- LDA(dtm, k = 10) # find 5 topic
> term <- terms(lda, 5) # first 7 terms of every topic
> (term <- apply(term, MARGIN = 2, paste, collapse = ", "))
                                       Topic 1                                   Topic 2
               "iphone, make, can, one, twitter"        "iphone, want, google, phone, better"
                                       Topic 3                                   Topic 4
"iphone8, android, facebook, iphone, samsung"        "new, iphones, iphone, iphone8, emoji"
                                       Topic 5                                   Topic 6
                 "get, now, new, service, free"           "store, iphone, app, phone, will"
                                       Topic 7                                   Topic 8
               "iphone, new, says, back, going"           "iphone, new, ios, look, battery"
                                       Topic 9                                  Topic 10
 "ipad, itunes, ipod, ipodplayerpromo, itune"     "iphone, hate, fingerprint, give, price"
> 
```



```
> lda <- LDA(dtm, k = 5) # find 5 topic
> term <- terms(lda, 5) # first 7 terms of every topic
> (term <- apply(term, MARGIN = 2, paste, collapse = ", "))
                                     Topic 1                              Topic 2
      "app, google, microsoft, new, iphone"       "iphone8, iphone, get, now, new"
                                     Topic 3                              Topic 4
        "iphone, phone, just, can, twitter"     "iphone, new, will, thanks, tech"
                                     Topic 5
"ipad, itunes, ipod, ipodplayerpromo, new"
```

qplot (date, ..count.., data=topics, geom ="density", fill= term[topic], position="stack")

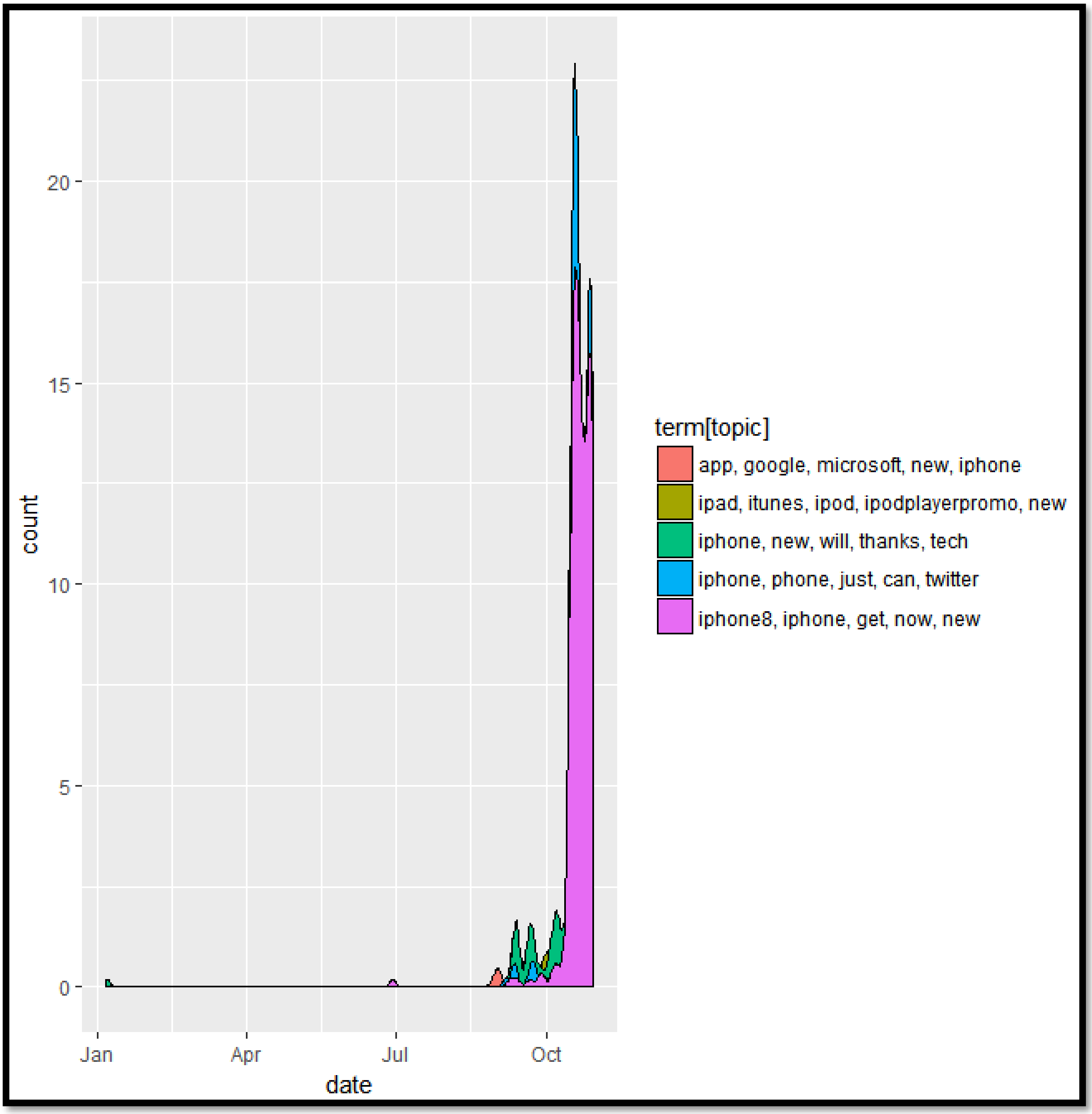
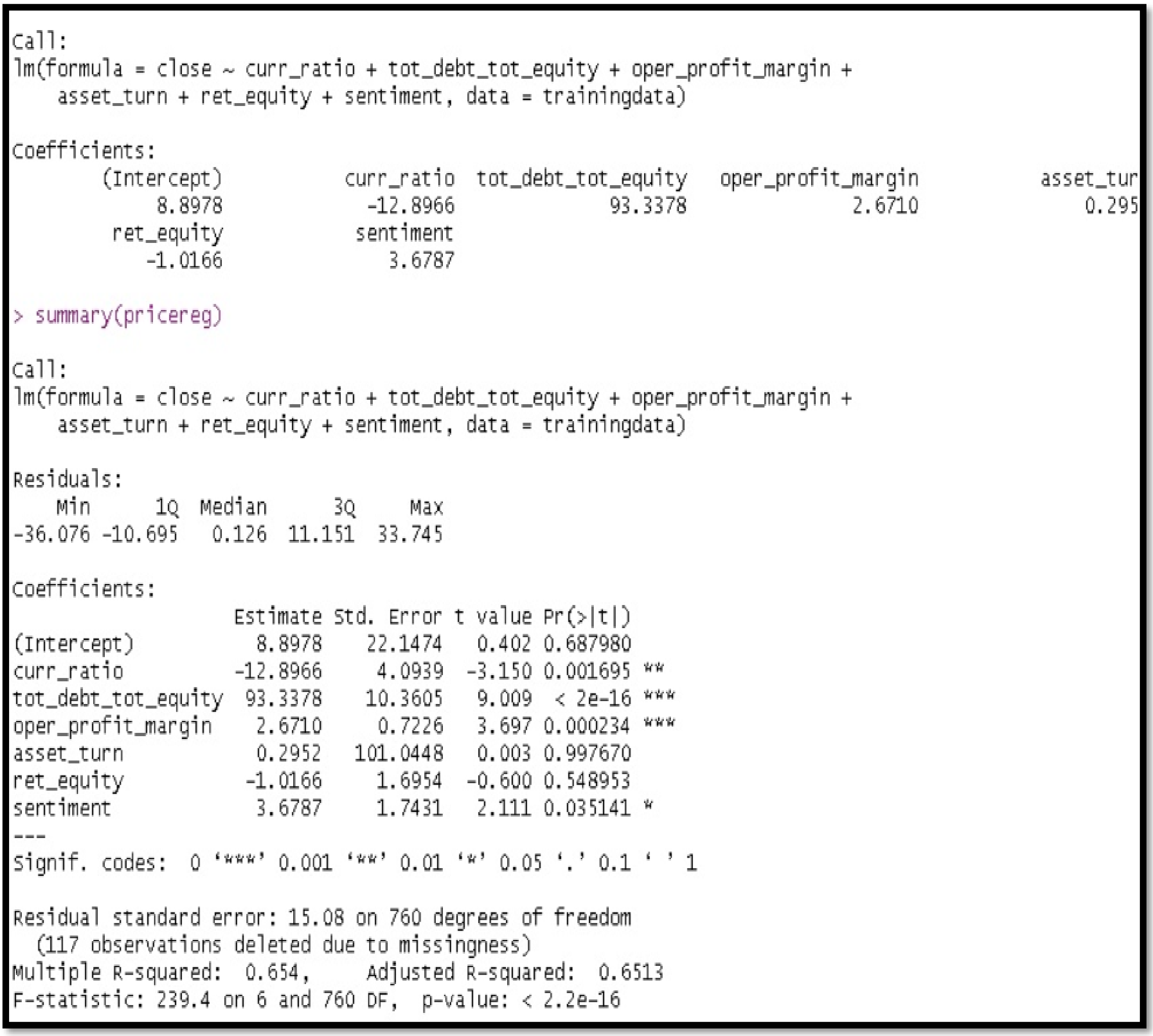


Fig 29: Topic modelling

Predictive Analytics

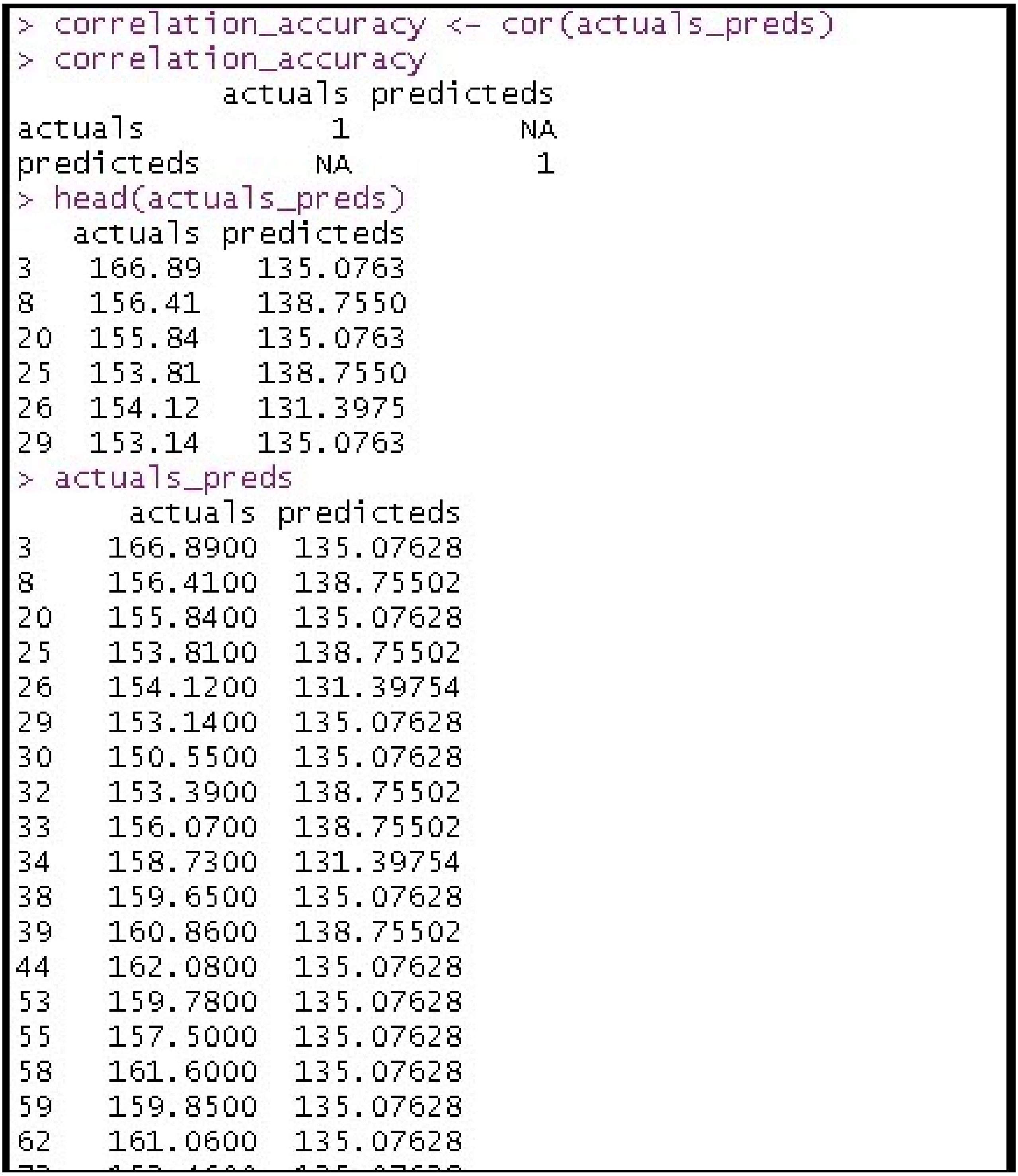
In order to find a suitable technique for predicting the stock price accurately, we tried a pool of 14 prediction models (7 with regression and other 7 for Neural Networks)

Regression with Sentiment scores couldn't result a reasonable prediction model, here are the few sample screenshots.



```
Call:
lm(formula = close ~ curr_ratio + tot_debt_tot_equity + oper_profit_margin +
    asset_turn + ret_equity + sentiment, data = trainingdata)

Coefficients:
        (Intercept)           curr_ratio  tot_debt_tot_equity   oper_profit_margin           asset_tur
             8.8978             -12.8966              93.3378               2.6710               0.295
         ret_equity            sentiment
            -1.0166               3.6787

> summary(pricereg)

Call:
lm(formula = close ~ curr_ratio + tot_debt_tot_equity + oper_profit_margin +
    asset_turn + ret_equity + sentiment, data = trainingdata)

Residuals:
    Min      1Q  Median      3Q     Max
-36.076 -10.695   0.126  11.151  33.745

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.8978    22.1474   0.402 0.687980
curr_ratio          -12.8966     4.0939  -3.150 0.001695 **
tot_debt_tot_equity  93.3378    10.3605   9.009  < 2e-16 ***
oper_profit_margin    2.6710     0.7226   3.697 0.000234 ***
asset_turn            0.2952   101.0448   0.003 0.997670
ret_equity           -1.0166     1.6954  -0.600 0.548953
sentiment             3.6787     1.7431   2.111 0.035141 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.08 on 760 degrees of freedom
  (117 observations deleted due to missingness)
Multiple R-squared:  0.654,     Adjusted R-squared:  0.6513
F-statistic: 239.4 on 6 and 760 DF,  p-value: < 2.2e-16
```

Fig 30: Results from neural network model



```
> correlation_accuracy <- cor(actuals_preds)
> correlation_accuracy
            actuals predicteds
actuals           1         NA
predicteds       NA          1
> head(actuals_preds)
   actuals predicteds
3   166.89   135.0763
8   156.41   138.7550
20  155.84   135.0763
25  153.81   138.7550
26  154.12   131.3975
29  153.14   135.0763
> actuals_preds
      actuals predicteds
3    166.8900  135.07628
8    156.4100  138.75502
20   155.8400  135.07628
25   153.8100  138.75502
26   154.1200  131.39754
29   153.1400  135.07628
30   150.5500  135.07628
32   153.3900  138.75502
33   156.0700  138.75502
34   158.7300  131.39754
38   159.6500  135.07628
39   160.8600  138.75502
44   162.0800  135.07628
53   159.7800  135.07628
55   157.5000  135.07628
58   161.6000  135.07628
59   159.8500  135.07628
62   161.0600  135.07628
```

Fig 31: Predictions from model

Neural Network model yielded a much better prediction model with the below sample outputs.

Neural Model:

We have multi-layer model having two LSTM layers, followed by one feed forward layer and one softmax based output layer.

Layer (type)	Output Shape	Param #
lstm_3 (LSTM)	(None, 5, 128)	69120
lstm_4 (LSTM)	(None, 128)	131584
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 64)	8256
dense_4 (Dense)	(None, 3)	195

Table 2: Multi-layer model output

Total parameters: 209,155

Trainable parameters: 209,155

Non-trainable parameters: 0

Neural model classifies Stocks based on 5 days of sequence and predicting 5th day. It's multi feature sequenc e model.

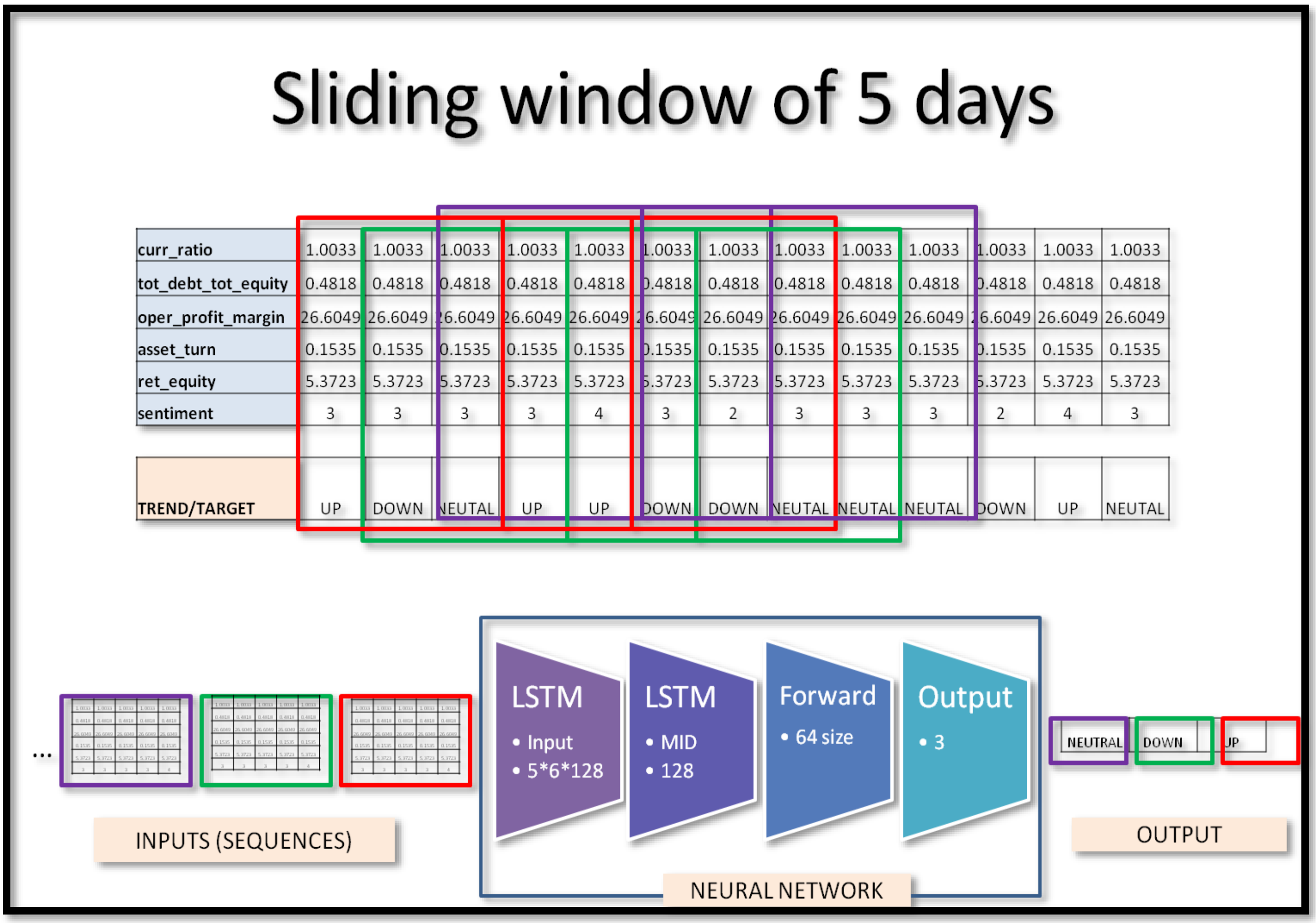
We have 3 classes :

1. Positive recommendation (where a day gain is more than 1%),
2. Negative recommendation (where a day loss is greater than 1%) and
3. Neutral recommendation for all other cases

Key features:

1. Inputs to ML algorithm (neural network) are formed with 5 days window to understand weekly impact.
2. 3 class neural network to predict: SELL/BUY/HOLD is done
3. Neural network has 3 layers, 2 LSTM to capture the sequence impact, one feed forward and then finally 3 class output classifiers.

4. Regression based model because of limited data is not considered, but same model can be converted to regression to predict prices itself.



Sliding window of 5 days

curr_ratio	1.0033	1.0033	1.0033	1.0033	1.0033	1.0033	1.0033	1.0033	1.0033	1.0033	1.0033	1.0033	1.0033
tot_debt_tot_equity	0.4818	0.4818	0.4818	0.4818	0.4818	0.4818	0.4818	0.4818	0.4818	0.4818	0.4818	0.4818	0.4818
oper_profit_margin	26.6049	26.6049	26.6049	26.6049	26.6049	26.6049	26.6049	26.6049	26.6049	26.6049	26.6049	26.6049	26.6049
asset_turn	0.1535	0.1535	0.1535	0.1535	0.1535	0.1535	0.1535	0.1535	0.1535	0.1535	0.1535	0.1535	0.1535
ret_equity	5.3723	5.3723	5.3723	5.3723	5.3723	5.3723	5.3723	5.3723	5.3723	5.3723	5.3723	5.3723	5.3723
sentiment	3	3	3	3	4	3	2	3	3	3	2	4	3
TREND/TARGET	UP	DOWN	NEUTAL	UP	UP	DOWN	DOWN	NEUTAL	NEUTAL	NEUTAL	DOWN	UP	NEUTAL

Fig 32: Sliding window of 5 days

Given model gives **96% of accuracy** in predictions. Out-sampling cases need to be handled better. Lag of data can help in better accuracy across classes.

Chapter 5

Recommendations

In this study, we have run predictive analysis on weekly returns of some selected stocks, based on a set of some fundamental financial ratios as well as sentimental data. While the neural network has shown better accuracy, the regression model has also acknowledged the sentiment data as a statistically significant predictor of stock price movements. The neural network model provides a rating of up, down or neutral which can be translated as Buy, Sell or Hold rating in the stock price prediction terminology.

Conclusions

Every investor has their own strategy, such as particular fundamentals they tend to be fond of and level of risk they are willing to accept. These types of analysis alone are becoming outdated and more effective investing tools have become evident to lending a helping hand in increasing portfolio performance. Machine-based algorithms, which can analyze many stocks simultaneously and determine quantifiable founded objective predictions, are becoming increasingly more popular to investors as an improved strategy for optimizing returns and mitigating prospective risk. There still is no flawless way of picking stocks but utilizing advanced algorithms based on predictive analytics in conjunction with reasonable risk management and fundamental analysis can potentially help improve portfolio performance for investors.

We have shown that a reasonably strong correlation exists between rise/fall in stock prices of a company with the trend in public opinions about that company expressed on social media, such as Twitter.

We have developed a sentiment analysis tool that can judge the type of sentiment present in the tweet. The tweets are classified into three categories: positive, negative and neutral. At the beginning, we claimed that positive emotions or sentiment of public in twitter about a company would reflect in its stock price. Our speculation is well supported by the results we have achieved.

Reference

Please find the details of the codes, dataset and reports that has gone into making this project at the below link:

https://github.com/sahilmakkar1983/GLCap