Sentiment Analysis for Stock Price Flow Analysis

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

In

Computer Science and Engineering

By

Nishant Suman (131221)

Under the supervision of

Dr. Pradeep Kumar Gupta

to

Department of Computer Science & Engineering
Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

`

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report entitled **"Sentiment Analysis for Stock Price Flow Analysis"** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from July 2016 to May 2017 under the supervision of **Dr. Pradeep Kumar Gupta** Assistant Professor, Department of Computer Science And Engineering. The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)
Nishant Suman(131221)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature)
Dr. Pradeep Kumar Gupta
Assistant Professor
Department of Computer Science And Engineering
Dated : 25/04/2017

`

ACKNOWLEGEMENT

We are grateful and indebted to Dr. Pradeep Kumar Gupta, Assistant professor, Department of Computer Science And Engineering for his help and advice in completion of this project report. We also express our deep sense of gratitude and appreciation to our guide for his constant supervision, inspiration and encouragement right from the beginning of this Semester report. We also want to thank our parents and friends for their immense support and confidence upon us. We deem it a pleasant duty to place on record our sincere and heartfelt gratitude to our project guide for his long sightedness, wisdom and co-operation which helped us in tackling crucial aspects of the project in a very logical and practical way.

Nishant Suman
(131221)
Computer Science and Engineering

`

Table of Contents

`

`

List of Abbreviations

1. DJI---------- Dow Jones Index
2. CART----------Classification and Regression Tree
3. kNN----------k- Nearest Neighbours
4. NB---------- Naive Bayes
5. SVM----------Support vector Machine
6. DFD---------- Data Flow Diagram
7. NLTK---------- Natural Language Toolkit
8. NLP---------- Natural Language Processing
9. AAPL ---------- Apple Inc Stock Symbol

`

LIST OF FIGURES

`

ABSTRACT

Social media comments and news on Internet regarding any company can impact the their flow of stock prices. Often, if a company has bad review about their product in social media can have impact on their stock prices a lot.

In this project, we investigate the relationship between StockTwits messages relationship with stock market movement. Specifically, we wish to see if, and how well, sentiment extracted from these feeds can be related to the shifts in stock prices. For this case we chose Apple Inc to perform the analysis. To answer this question, we construct a model, estimate its accuracy, and put it to the test on real market data.

The state of the art in sentiment analysis suggests there are 2 important mood states that enable the prediction of mood in the general public. The prediction of mood uses the sentiment word lists obtained in various sources where general state of mood can be found using such word list or emotion tokens. With the number of messages posted on StockTwits, it is believed that the general state of mood can be predicted with certain statistical significance.

`

1. INTRODUCTION

Social media comments and news on Internet regarding any company can impact the their flow of stock prices. Often, if a company has bad review about their product in social media can have impact on their stock prices a lot. There are many platforms in Internet where user can share their opinion about any thing such as Twitter, Facebook, StockTwits etc. The posts in these platform related to a company can have an impact on the stock price of that company. In India BSE(Bombay Stock Exchange) is used to trade shares, In America DJIA Index is used for the same.

StockTwits

StockTwits, as shown in figure 1.1, is a micro-blogging platform where a user can post or read messages of up-to 140 characters. Registered user can read or post messages while the unregistered user are only able to read messages. It is mainly used by investors, traders and entrepreneurs which help them to share ideas among them. It is similar to Twitter but it uses cash tags($) to tag a company or market. StockTwits was founded in 2009 by Howard Lindzon and Soren Macbeth. Its main headquarter is situated in San Francisco and it has more than 25 small headquarters in more than 25 places. In March 2016, StockTwits has more than 300 thousand active users and more than 340 thousand messages were posted in each day. It has also received Shorty Award in 2008 in finance category.

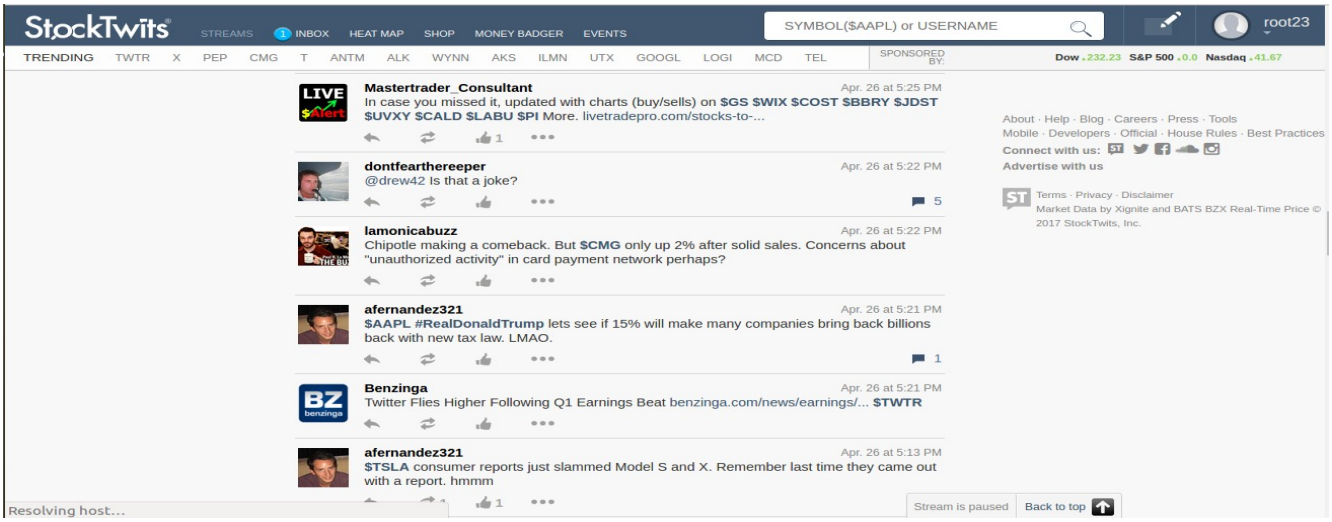

Figure 1.1: StockTwits Platform

`

Stock Market

A stock market is the collection of stocks or shares of various companies where a client is able to buy or sell stock or shares through buyers or sellers. Stock market is differentiated in many ways. One common way to differentiate between stock market is by the country in which it is located.

Stock Exchange

A place or organization where stock trader can buy or sell stocks is called as Stock Exchange. To go public, company has to enlist itself in stock exchange. Stocks can be traded through a dealer.

Sentiment Analysis

Sentiment Analysis is a process of finding the mood of a sentence. The mood can be positive, negetive or neutral. It is also known as opinion mining. It is used to derive opinion or sentiment of speaker. Generally it is used to determine how a person feel about a particular subject or topic.

Apple Inc

Apple Inc is an American based multinational company which design, develops and sell their mobile and computer products worldwide. Its headquarter is located in Cupertino, California. It was founded in April of 1976 by Steve Jobs, Steve Wozniak and Ronald Wayne. It is traded as NASDAQ:AAPL, as shown in figure 1.2, and has joined the DJIA (Dow Jones Industrial Average) for its stock market trading.

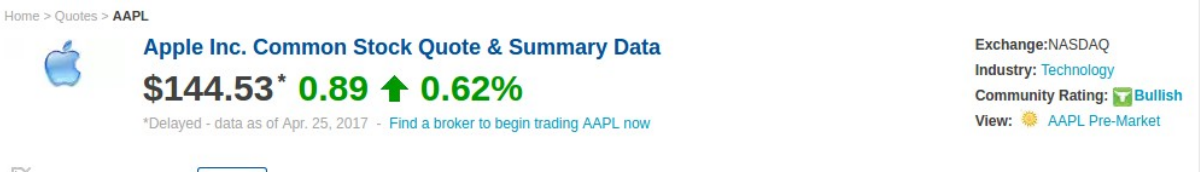

Figure 1.2: Apple Inc Stock Quote

`

1.1 Problem Statement

Consistently, there are a lot of posts that are shared online through any social media like twitter or StockTwits. It contains posts which may be considered as important data about every theme. Through StockTwits alone, more than 400 Thousand messages are posted for every day. It has at-most 140 characters, that is more than 56 million characters produced every day. Despite the fact that every message may not appear to be greatly significant, it has been used to collect a large amount of message that can provide profitable knowledge about open state of mind and assessment on certain fields.

In this project, we use Naive Bayes classification technique of machine learning that are generally use to classify the sentiment out of any sentence. We also intend to find the degree to which messages are associated to stock costs on a day based scale and monthly based scale, and also inquire about which singular words from messages are connected with changes in stock costs. Generally, stock value connections include numerous more variables, however we will simply take a look at the connection between messages and variations on day based scale.

1.2 Aims and Objective

The overall aim is to analyze the StockTwits's messages Sentiment posted by the users regarding Stock Market and analyse the flow of stock market with respect to social media comments.

1.3 Methodology

In this project we have calculated the StockTwits messages Sentiments and use it to analyse the shift in stock market. Our prime methodology is to get streaming data from various users, news blog etc using its API and use that streaming data to get the sentiments of user on various companies. Then we extracted the stock data of Apple Inc in day to day basis. After that we visualise both of them to analyse the flow of stock market with respect to the social media comments on Apple Inc.

`

1.4 Organisation

In Chapter 1 we have discussed about basic concept of Sentiment Analysis, Stock Market and Twitter. We also provide our aim in this project and a brief description on how we are going to achieve our goal.

In Chapter 2 we would be providing with the basic terminology about the different research paper read by us. We would be providing with facts and figures about different concepts we studied in those research papers.

In Chapter 3 we are going to provide a model of how the project is done on the basis of developments:-

- Analytical
- Experimental
- Statistical

In Chapter 4 we have given a proper analysis of our sentiment analysis engine along with its accuracy.

In Chapter 5 we have given the conclusion about our project.

`

2. LITERATURE SURVEY

In the paper published by Sang Chung and Sandy Liu[1], at 2011, on anlysis of stock market based by analysing the sentiments of tweets posted on twitter, they found out that the bullish beahvior of the market can be predicted by the twitter sentiments but, the bearish beahvior of market doesnot seem to follow the sentiments extracted from those tweets. They provide a possible reason for that the did not have enoough tweets due to the Twitter's privacy policy. They came out with a graph as shown in figure 2.1a and 2.1b. Also users can tweets all day while stock market is opened for only limited hours. They came out with the following twitter sentiments graph and stock price fluctuation:

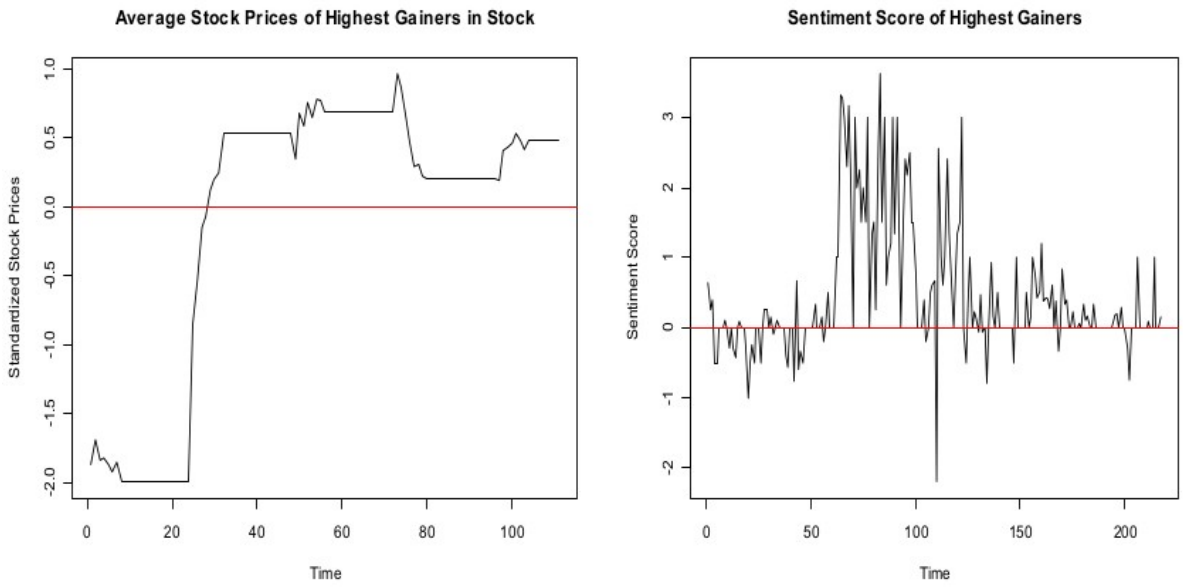

Figure 2.1a shows Deviation of stock prices with respect to the price os stock as Market opens and Figure 2.1b shows the sentiments fluctuation of twitter user with respect to a stock.

In the paper published by Linho Zhang[2], at April 2016, he tested various classifiaction techniques to classify the sentiments of the tweets. He mainly used Naive Bayes Classification, Support Vector Machine and Maximum Entropy Classification. He discovered that among those three classifiers, Support Vector Machine provides the most accurate result by cross validatin them, but that accuracy was also wasnot high enough. The main challenge faced by him is to drag out proper meaning for the words such as, 'ur' which is subsituted by people for 'your' etc.

`

In the paper published by Nuno Oliviera[4], she studied the relation between the posts in StockTwits and stock market. She performed her analysis for six major companies by measuring their posting volume in StockTwits and their trading volume in stock market. She discovered that using microblogging data, predicting the stock market variables such as returns and volatility is much more harder task.

Social Media comments can affect the flow of market and may use to study the behavior of market as suggested by SA Bogle and WD Potter[7]. In their paper, they analysed the Twitter posts' sentiment with Jamaica Stock Exchange and found out that they can predict the flow of market by 87% and their correlation coefficient for price prediction is 0.99. Their paper indicates that the flow of market is not only dependent upon traditional quantitative input but also depends on qualitative inputs posted on social media platforms.

In the paper by Anshul Mittal and Arpit Goel[22], they try to find out about the relationship between 'pucblic sentiment' and 'market sentiment'. In their work, they try predict moods of the public by using previous day's market flow of DJIA. They use self organising fuzzy neural network to predict the movement of stock market movement with respect to twitter feed. Their efficiency of model for predicting the stock market is 75.56%. They find out among all their moods, only calmness and happiness can predict the actual flow of DJIA values. They main drawback in their work is they try to predict the movement of stock market only using the twitter feeds of enlish speaking people not all public opinions.

2.1 Aspect-based Sentiment Analysis

As shown in Liu B.'s Paper[19], anyone who want to analyse sentiment of any archive, he could classify it in three ways. That concludes that if any tweet posted in the twiiter can be inspected in following ways:

- Document level, whether the whole message has positive or negative emotion?
- Sentence level, whether each sentence of message have positive or negative emotion?
- *Entity and feature/aspect level*

`

Document level (Message Level)

The most easiest of the above mentioned three aspect is finding out the positive or negative emotion of any document (In this case message) is anlysing its emotion in document level. It does not take it into account that whether the document is different or not. A mentioned by Antoniou in his research paper of 2012, document level could provide satisfied result to the user who want mine the sentiment out of any document (message).

Sentence level

As mentioned by Zhang, Xu and Li(2012)[41] and Hu and Liu(2004)[19], there are many ways to find the sentiment at sentence level that is whether a particular sentence has positive or negative sentiment. The sentence may contain different word that points to different emotion of the user. The opinion of any sentence is a sentence which contain more than one sentiment or opinion. Analysing sentiment at sentence level is entirely different working as conpared to the entity level.

Entity and feature/aspect level

Analyzing a message at the feature/aspect level is the best way to find the sentiment of any subject and it gives more in depth sentiment analysis. It can provide sentiment analysis to understand different view of any message such as reviewing a product like iphone 7 or comparing products of same field like iphone 7 and pixel. Suppose there is a message about Google's pixel (an android based smartphone by Google) that only have review about it. The user may give positive or negative sentiment about the product but in the mean time another user give negative review about it may about phone's design or its processor. The aftereffect of a consistent estimation classifier would be a characterization as unbiased or a score close to zero (if the conceivable scores can likewise get negative qualities) for the entire tweet. Along these lines, however a vital bit of data gets lost.

For this situation a there may be a post that may state that iPhone 5S has the best battery ever, while HTC One S has the best show. Grouping this specific tweet just as positive (since it contains two circumstances "best") would prompt to deluding conclusions.

`

Separating these elements and viewpoints is a great deal harder than basically deciding algorithms and the and seperating the mixtire of these two (extricating elements and deciding sentiments) is considerably harder. One answer for the issue is the organization of Ontology-based methods (Kontopoulos, Berberidis, Dergiades, and Bassiliades, 2013)[21]. It comprises of terms (e.g. Cell phone), connections between terms, properties, confinements and different explanations. Along these lines it gives an entire vocabulary to show effectively any area. Ontologies constitute the primary segment of Semantic Web.

The philosophy that is followed in (Kontopoulos, Berberidis, Dergiades, and Bassiliades, 2013)[21] is isolated in two particular stages. Above all else, the area metaphysics that will be utilized is constructed and after that estimation examination is performed in tweets in view of this specific philosophy. The ontologies were fabricated utilizing the Formal Concept Analysis (Ganter and Wille, 1999)[16], which applies a client driven well ordered system for making area models. The greatest preferred standpoint of this approach is that the philosophy is assembled continuously as indicated by the necessities that are introduced in the tweets. In the case of the tweet discussing the properties of cell phones, this approach would not simply locate a prepared philosophy for cell phones with a great many traits and connections, yet it would make another one construct just in light of the given information set. Along these lines we wind up with a more information particular philosophy with littler size.

In another applicable paper (Nebhi, 2012)[29] the creator concentrates just in the data extraction from tweets. He utilizes a prepared philosophy from DBpedia and makes an incorporated disambiguation module in light of prevalence score and sentence structure based likeness with a specific end goal to choose every time the correct element from the given metaphysics. That is a current need, since basically separating words out of a tweet is not generally enough with a specific end goal to group it into the correct cosmology. There are many words (like "Washington" for instance) that may allude to more than 90 unique substances of the DBpedia philosophy.

2.2 Sentiment Analysis in Social Networks

As we have as of now observed, with Sentiment Analysis we attempt to concentrate feeling or state of mind out of an archive or a bit of it concerning the entire record or a particular part of

`

it. In informal communities a report could be a tweet or a post at a long range informal communication site like facebook or any bit of content conceivably containing some sentiment. In the event that somebody could, in some programmed way understand each conceivable bit of content in the web, he would be certainly the most effective individual on earth. Individuals impart insights and individual encounters about any conceivable subject in the web. They submit surveys, compose remarks about different items or individuals (e.g. lawmakers, competitors and so on.) or compose articles in sites, smaller scale online journals, discussions and different types of sites.

Performing sentiment mining from miniaturized scale sites is viewed as less demanding contrasted with different sorts of archives. The primary explanation behind this is the short size of archives utilized as a part of smaller scale blogging, driving those the vast majority of the circumstances to be straighter to the point. The most celebrated small scale blogging administration right now is Twitter and that is the reason that we will utilize it as a source of perspective the vast majority of the circumstances. Besides a large portion of the exploration that has been done on Sentiment Analysis in smaller scale writes up to now is around Twitter.

2.3 CityFalcon

CityFalcon, as shown in figure 2.2, is a platform which collets data from more than 200 finanancial sources (including StockTwits, Twitter etc) and shows relevant financial news and information to all investors based on their choice. It scans all major internet resource in order to collect relevant finacial data to its users. Along with that, it also collects data and do sentiment analysis on them to provide their percentage of bullish of bearish behavior with respect to market. It is a startup which is launced in July, 2014. Users can filter their watchlist about the company in order to gather data from it. In provide information in all languages so that people from all over the globe can access it and use its data for investment purpose.

`

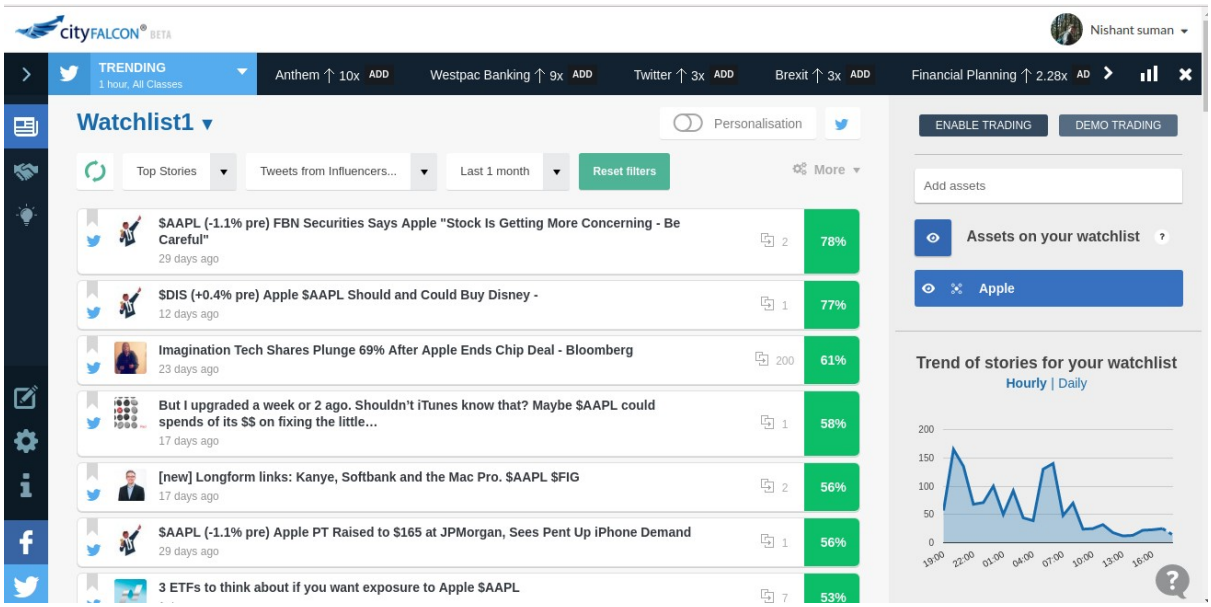

Figure 2.2: CityFalcon

`

3. SYSTEM DEVELOPMENT

In this section we have shown various UML diagram for our project and discuss about the methods and approaches which have been followed to analyse the StockTwits messages' relation with respect to the flow of the market.

3.1 UML Diagram

UML Diagram of Unified Modelling Language Diagram is used provide a static stucture for the overall process which have been used to create a software or analysis.

3.1.1 DFD

DFD or Data Flow Diagram is a way of representation graphically in order to provide the basic flow of data in a software model. The box stucture shows the input/output of data and the circular structure shows the process of the data. It is mapped using the arrow lines wich represent the flow of data from one process to another process. It can be elaborated in many levels (genrally 4 or 5). We have provided the four basic levels of DFD, as shown in figure 3.1, 3.2, 3.3, 3.4, which shows the flow of data relevant to our project.

Level 0

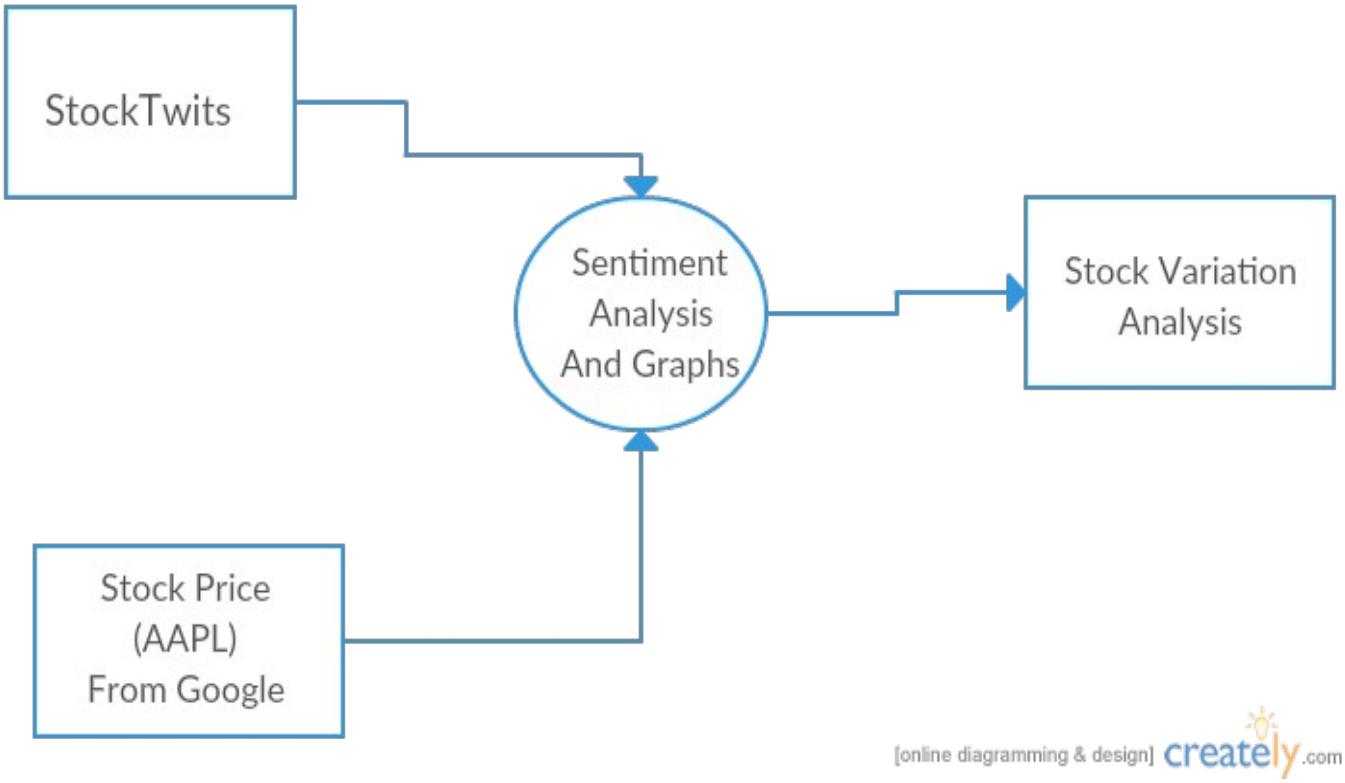

Figure 3.1: DFD level 0

`

Level 1

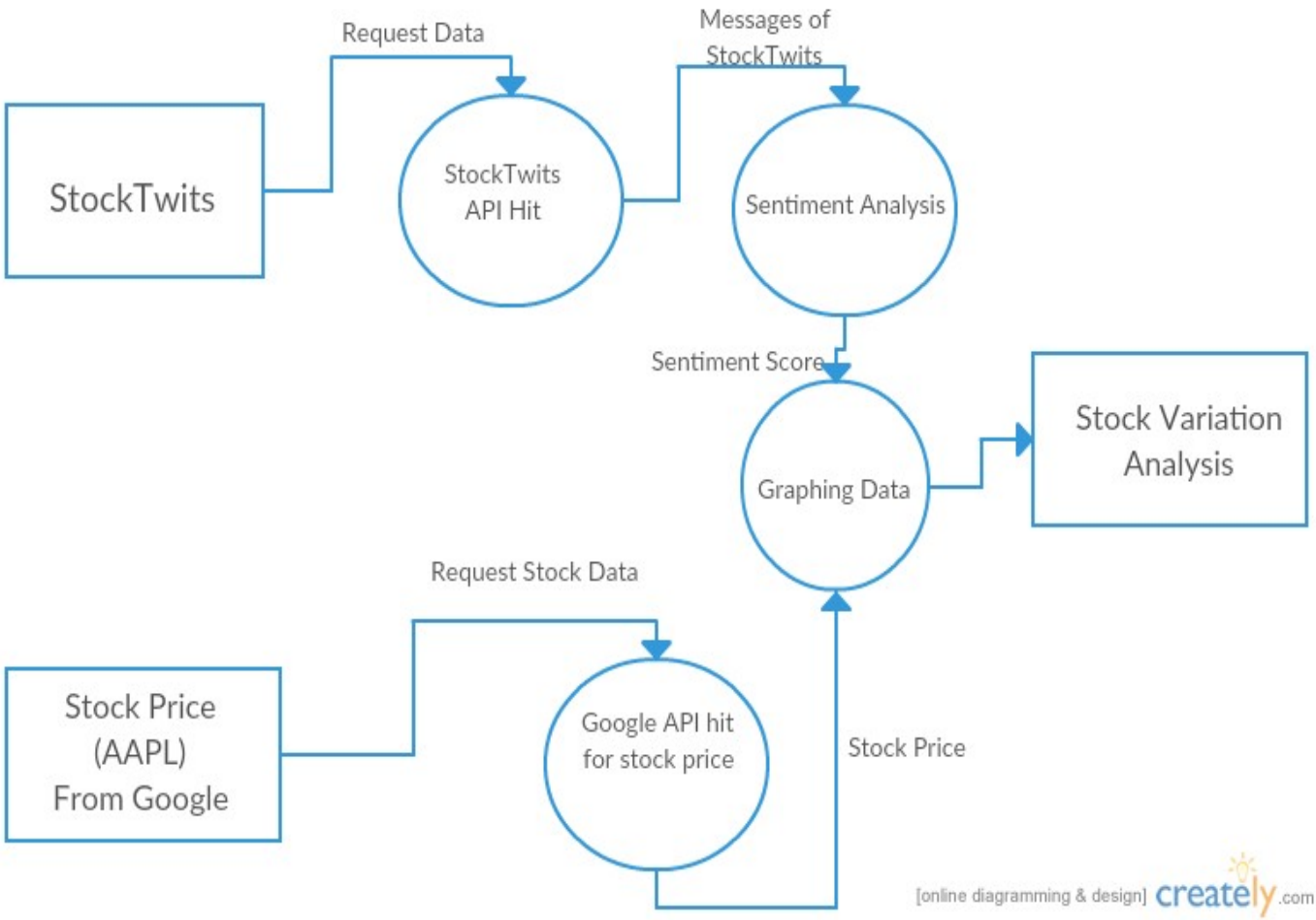

Figure 3.2: DFD level 1

`

Level 2

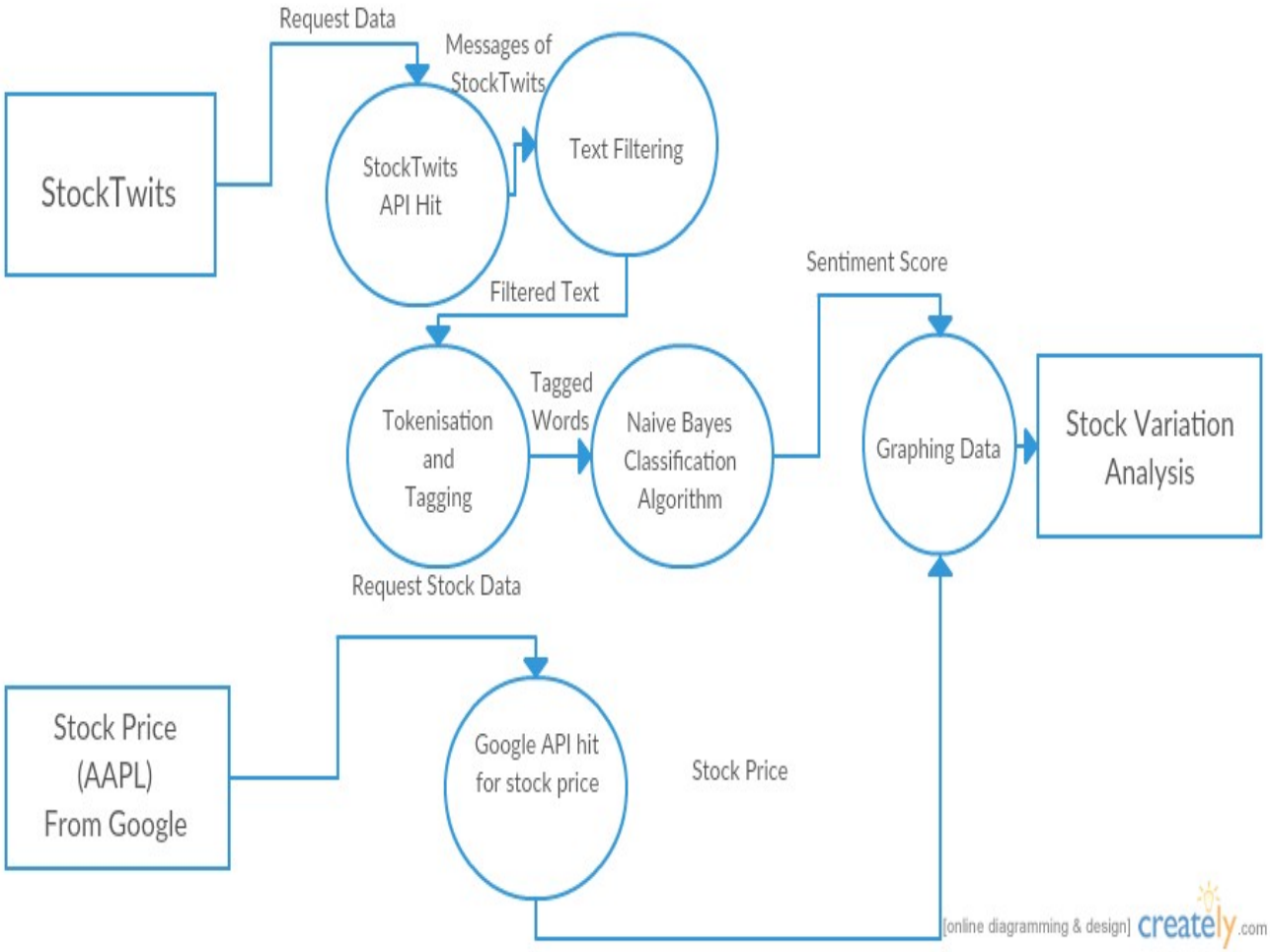

Figure 3.3: DFD level 2

`

Level 3

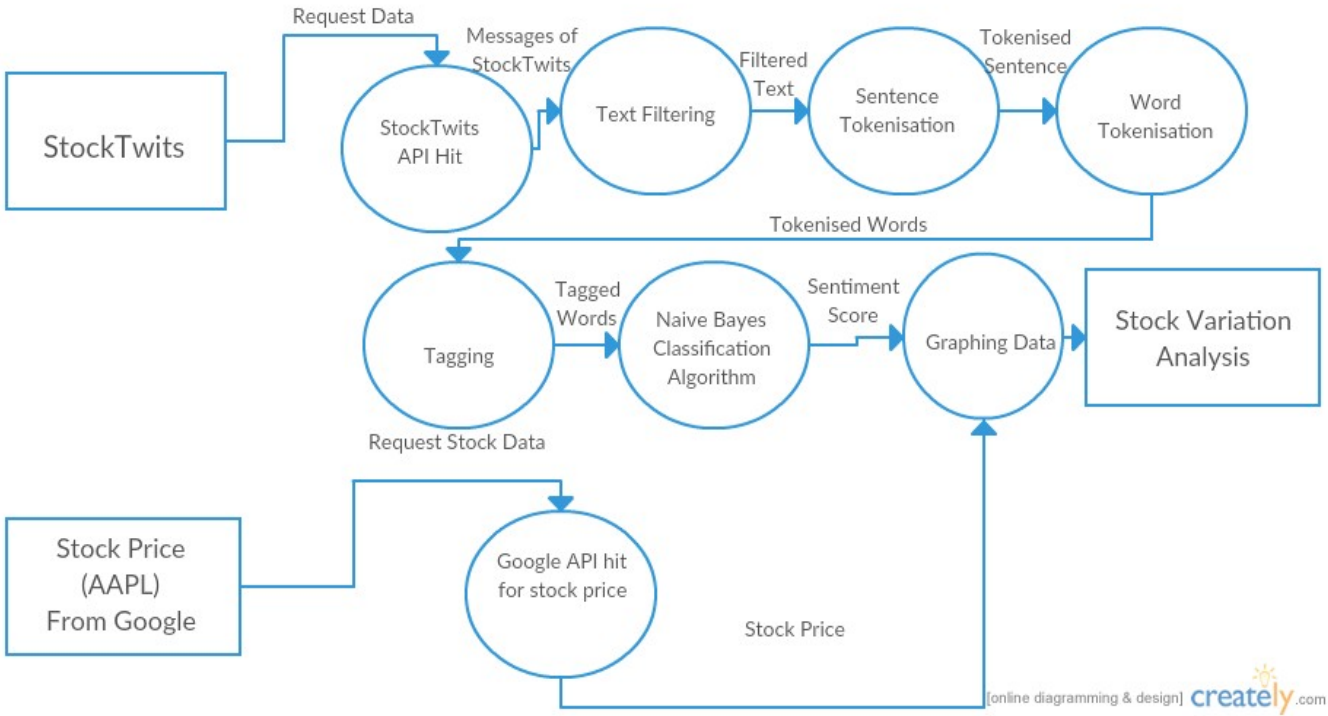

Figure 3.4: DFD level 3

`

3.1.2 Architecture

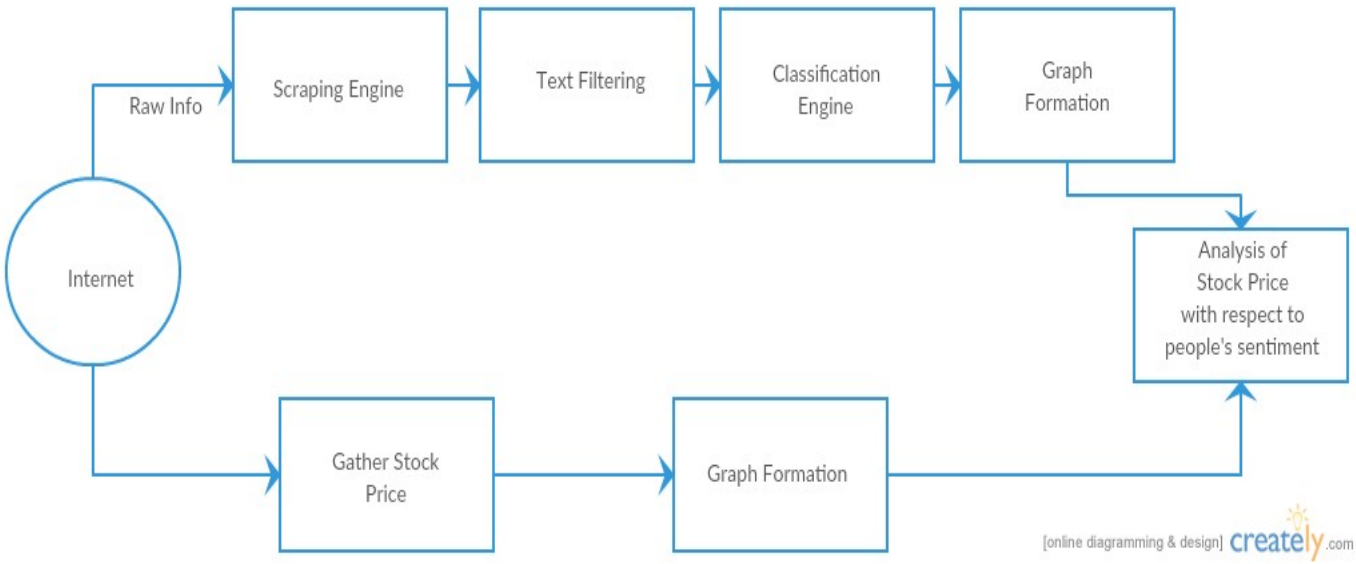

Figure 3.5: Complete Architecture of Project

In the architecture, as shown in figure 3.5, at first data from Internet(StockTwits) is scraped using scraping engine(along with StockTwits API). The data scraped using stock twits is in the raw form(JSON). Now the scraped data contains many unnecessary data which is not being used in the project, that is, we need only data which is relevant to us. Also the relevant data also contains links, mentions, tags and emoji/smiley which is unnecessary for us, so for that text filtering is done. Now, the filtered text is passed in to classification engine which classify the nature of data, that is, whether the data show bullish behavior or bearish behavior. The classification engine performs tokenisation, removal of stop words, tagging and finally Naive Bayes classifier to classify the data. After classifying the data, sentiment score of that data is calculated using its most informative features. Also stock market data is also gathered from the Internet(Google's API). These stock market data and sentiment scores are plotted in the graph. Finally analysis of variation of stock market with respect to people's sentiment is done.

`

3.2 Requirements

It contains functional as well as non functional requirements of the project.

3.2.1 Functional Requirements

1. The user should be able to run any classifier to further train any model so that they can improve the performance of classifier.
2. The system should be able to run for long time without any problems as the processing for such large volume of data takes time.
3. Efficient system should be used so that it consume little storage and it must have hi processing power so that it will take less time to run.
4. The model that is used to train using the data provided should be reliable even is there is any error in textual form.
5. The user should be able to use same mechanism for training and testing of data.

3.2.2 Non Functional Requirements

1. The system should be able to respond in reasonable time that is the system must be fast enough to run any classifier.
2. It must be easy to use.

3.3 Gathering messages from StockTwits

3.3.1 Accessing StockTwits messages

To access StockTwits' messages, we have written a script which intract with StockTwits using its API. At first we need to register out app in StockTwits API in order to access messages. From their we get consumer key and consumer secret (not changeable) which would enable us to access messages. We also need to generate access token and access secret (changeable) which along with consumer key and consumer secret will help us to gather messages for our project. These keys and secrets should be kept in secret as it will enable anyone to access the StockTwits account. We can only access messages of those users whom we follow on StockTwits.

`

The consumer key, consumer secret, access key and access secret are in the form of alphanumeric keys shown below:
Consumer Key (API Key):Up0sS17oKtNetiM2dFAQxWG
Consumer Secret (API Secret):
6SQHHOn8IOrFhUYSQCLi0lhpuSXVJiWsX676pAcL7Wk7bGB
Access Token:1329898987-yGbTmcpdz1LiKnwQ4Mc4QoSpsrBUJ8jOKGA
Access Token Secret:G7aOtbrnlc0Kw6ccNiMPebl9MLF2kO6UvD5Kqyma

To access data from StockTwits we need to hit its api using this base url “https://api.stocktwits.com/api/2/” and base parameter which is a secret key in order to make connection with stockTwits. Then we used this link “https://api.stocktwits.com/api/2/streams/friends.json” to access our dashboard in order to get messages posted by the people. The use get(url, params, timeout, stream) function of requests library finally get the dashboard feed.
Since we need to keep our connection open in order to gather all upcoming tweets about the companies, we have used to used streaming API of StockTwits for which we run the infinite loop which activates in every fifteen minutes to get data continuously. We have tracked the data using cashtags($) provided by the StockTwits. The api returns the data in json format which we parse it to get the particular property of it which contains messages by the users.
Using the above mentioned techniques we have gathered a lot of messages. Some of those messgaes are:
Long stock alternative in $AAPL for bulls helps offset potential near term slippage on entry after 5% rally in week.
Apple's $AAPL latest patent suggest the firm is building a self-driving$AAPL $Apple #Jobsreport #FeelgoodFriday & They like to have Market close aLiL Green $AAPL

`

3.3.2 Messages Filtering

The message collected contains many irrelevant information which is not related to Apple INC. These irrelevant information is need to be removed. So, any sentence which doesnot contains the word Apple, Apple Inc, $Apple, $AAPL, AAPL, $AppleInc or $AppleINC is removed.

We need to filter the collected messgaes i.e. make it free from dollar tags, hyperlinks and smiley. In order to do that we have used regular expression. Following regular expressions are used to filter the messages:

<[^>]+>', # HTML tags
?:@\w_]+)', # @-mentions
(?:\#+[\w_]+[\w\'_\-]*[\w_]+)", # hash-tags
http[s]?://(?:[a-z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-f][0-9a-f]))+', # URLs
?:(?:\d+,?)+(?:\.?\d+)?)', # numbers
(?:[a-z][a-z'\-_]+[a-z])", # words with - and '
(?:[\w_]+)', # other words
(?:\S)' # anything else

We also divide the messages manually positive messages and negative messages and also used the underlying corpora of NLTK for sentiment analysis into in order to provide learning data for the sentiment analysis engine.

3.4 Sentiment Analysis

In order to do sentiment analysis on the gathered messages we used NLTK library of python which a toolkit for Natural Language Processing(NLP) methodology.

- **Text Classification**

Text classification is used to classify the messages as positive or negative tweets which show bullish or bearish behavior of message with respect to stock market. Since positive and negative messages are already stored in separate documents, we need to open these

`

documents and shuffle the messages so the learning of machine is unbiased. If it is not done then there may be a chance that we train machine with all positive messages and some negative messages and test it against only negative messages or vice versa.

- **Tokenization**

Sentence Tokenization

Breaking a bunch of text in to a sentences is called sentence tokenization. Genarally a messages contains only one sentence but there is achance that a message may contain more than one sentence , so we need to split that message into many sentences. To do that using NLTK we used sent_tokenize() function provided by NLTK.

Word Tokenization

Breaking a sentence into words is called word tokenization. To perform that we use word_tokenize() funtion provided by NLTK and tag them as positive or negative word based on the file through which tweets have been taken.

- **Stop Words**

Generally a sentence contains word such as that, is, as etc which is not going to give any impact in sentiment analysis. These words are generally useless words or filler words. We need to get rid from these words present in the tweet. So after tokenizing sentence into words we need to remove these stop words from the data collected. So we have used the stopwords.words('english') function provided by the NLTK in order to do so. The major drawback of this function is that it cannot remove the sarcastic words or phrases.

Here are some the stop words provided by NLTK library:

'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the', 'themselves', 'until', 'below', 'are', 'we', 'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were', 'her', 'more', 'himself', 'this', 'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all', 'no', 'when', 'at', 'any', 'before', 'them', 'same', 'and', 'been', 'have', 'in', 'will', 'on', 'does', 'yourselves', etc.

`

- **Stemming**

There are many variation of a single word which has same meaning if tense is not involved. For example: I was riding a horse or I was taking a ride on horse. Here root word is ride and both sentence impose same meaning. So we need to find the root of each word which is done using stemming. Stemming is a method of finding the root words from different word. So to perform stemming, Porter Stemmer is used in this project. This stemmer is provided by the NlTK library of python.

- **Named Entity Recognition**

Named Entity Recognition one of the major form of chunking data in NLP. It is used to get the knowledge by the machine to get entities like places, things, location etc. It is then tagged with the word and stored in the feature set.

In figure 3.6, hierarchical named entity recognition a sentence is shown:

Figure 3.6: Named Entity Recognition of a StockTwits message

- **Tagging**

For classifiers to understand and learn the sentiment of word we tagged each word as pos or neg tag which show the bullish and bearish behavior of word in a particular sentence and in particular perspective.

- **Classifiers**

`

To get classifier for our sentiment analysis we have used scikit-learn library of python and used the basic Naive Bayes Classifier and compared this classifier against varoius other classifier in order to find the most efficient classifier for sentiment analysis.

Naive Bayes Classifier

It is a probabilistic classifier based on Bayes' hypothesis in probablity in statistical analysis. It is one of the least used classifier to see how precisely it functions and it is utilized a large portion of the circumstances.

It work fine on any dataset that can be appeared as arrangements of datasets. A dataset could be anything that can be available like StockTwits messages, tweets, facebook posts which is used in this project. On account of records or tweets when all is said in done, components can be the words. It expect that all factors are free with each other, implying that the nearness or nonappearance of a particular element is absolutely disconnected to that of some other element.

In Naive Bayes classifier, to find the probability of any data, it uses Bayes rule of probability. This Bayes rule is expressed as P(label | feaures) with respect to P(label) and P(features | labels).

There are two parameters to calculate sentiment using Naive Bayes Classifier:

1. P(label) = It is a probability received by each input label. There is no information about the input's feature is provided.

2. P(features | label) = It is a probability calculted based on that a given feature will receive a given value on each provided label for it.

Thus, P(label | features) = P(label)*P(features | label) / P(features).

Since all features “naive” assumed as independent with respect to each other,

So, P(label | features) = P(label)*P(f1 | label)*P(f2 | label)*.....*P(fn | label) / P(features)

It will consume a lot of time to compute P(features) explicitly, Naive Bayes Algorithm calculates each numerator based on their label and normalizes them as ,

`

P(label | features) = P(label) * P(f1 | label)* P(f2 | label)...P(fn | label) / sum(l)(P(l)*P(f1 | l)*..* P(fn | l))

here f1, f2, f3... fn are features and l is label

Naive Bayes Algorithm

classify(feature-sets):

```
    f ← copy(feature-sets)
    for fn in f.keys():
        if(l,fn) is in self.f_probdist:
            break
        else:
            del(feature-sets[fn])
    endfor
    for l in self.l:
        logprob[l] ← self.l
    endfor
    for l in self.l:
        for(fn, fv) in f.items():
            if(l, fn) in self.f_probdist:
                fprob ← self.f_probdist[l, fn]
                logprob[l] ← logprob[l] + f_prob.logprob(fv)
            else:
```

`

```
                    logprob[l] ← logprob[l] + sumLog([])
            endfor
        endfor
        return DictProbDist(logprob)


train(cls, l_fs, ELEpd):
        l_f ← freqDist()
        f_freqdist ← dict(freqDist)
        f_v ← dict(set)
        fn ← set()
        for fs, l in l_fs:
            l_f [l] incremented by 1
            for fn, fv in fs.item():
                f_freqdist[l, fn][fv] is incremented by 1
                f_v[fn].add(fv)
                fn.add(fn)
            endfor
        endfor
        for l in l_f:
            n_samples ← l_f[l]
            for each fn:
                c ← count(f_freqdist[l, fn])
                if n_samples – count > 0:
```

`

f_freqdist[l, fn] is incremented by (n_samples – count)

f_v[fn].add(none)

endfor

endfor

l_probdist ← ELEpd(l_f)

f_probdist ← {}

for(l, fn) and fd in f_freqdist.items():

probdist ← ELEpd(fd, bins ← len(fv[fname]))

f_probdist[l, fn] ← probdist

endfor

return cls(l_probdist, f_probdist)

- **Sentiment Score**

After classifying data based on their sentiments into bullish or bearish behavior, a sentiment score for each data is generated using the most informative features of data. Most informative feature for each word is the score of each word which will impact in the sentiment calculation of data. So there is a larger impact if sentence contains any word which has large score. Now, to calculate sentiment score, most informative feature of each word is gathered. If any word has negative sentiment its most informative feature is converted into a negative number. After gathering most informative feature for each data a sum of all most informative features is calculated and it is divided by total for most informative features to calculate sentiment score. The sentiment score thus varies from 0 to 1.

sentiment score = sum(mif(word)) / total(mif(word))

here mif is most informative feature

sum(mif(word)) is the sum of all informative feature keeping in mind that it is negative for negative sentiment.

`

total(mif(word)) sum of all informative feature disrespect to whether word has positive sentiment or negative sentiment.

3.5 Gathering Stock Market Data

To gather stock market data from Google finance(finance.google.com), this link 'http://finance.google.com/finance/info?client=ig&q='+symbol_list is used. In this symbol list contains the list of stock market symbols which in our case is AAPL (for Apple Inc stock exchange). The request() function of python is used to hit the link and obtain the data. The data is returned in the form of JSON structure from which stock values are obtained. This scraping of stock value is made to continuously run between the time of opening of market and closing of market. The returned JSON format for stock market is:

```
{
   'id'     : 'ID',
   't'      : 'StockSymbol',
   'e'      : 'Index',
   'l'      : 'LastTradePrice',
   'l_cur'  : 'LastTradeWithCurrency',
   'ltt'    : 'LastTradeTime',
   'lt_dts' : 'LastTradeDateTime',
   'lt'     : 'LastTradeDateTimeLong',
   'div'    : 'Dividend',
   'yld'    : 'Yield',
   's'      : 'LastTradeSize',
   'c'      : 'Change',
```

```
    'c'      : 'ChangePercent',
    'el'     : 'ExtHrsLastTradePrice',
    'el_cur' : 'ExtHrsLastTradeWithCurrency',
    'elt'    : 'ExtHrsLastTradeDateTimeLong',
    'ec'     : 'ExtHrsChange',
    'ecp'    : 'ExtHrsChangePercent',
    'pcls_fix': 'PreviousClosePrice'
}
```

Here 'l' contains the value of stock market.

`

3.6 Graphical Representation of Data

3.6.1 For Sentiments

For plotting the sentiments score of StockTwits messages, bar chart is used to plot the data. To do this, Matplotlib (a python's matlab library) is used. The sentiment's graph y-axis varies from 0 to 1 while the x-axis is varied from (0:00 hrs to 23:59 hrs). A sample of sentiment's score graph is shown in figure 3.7:

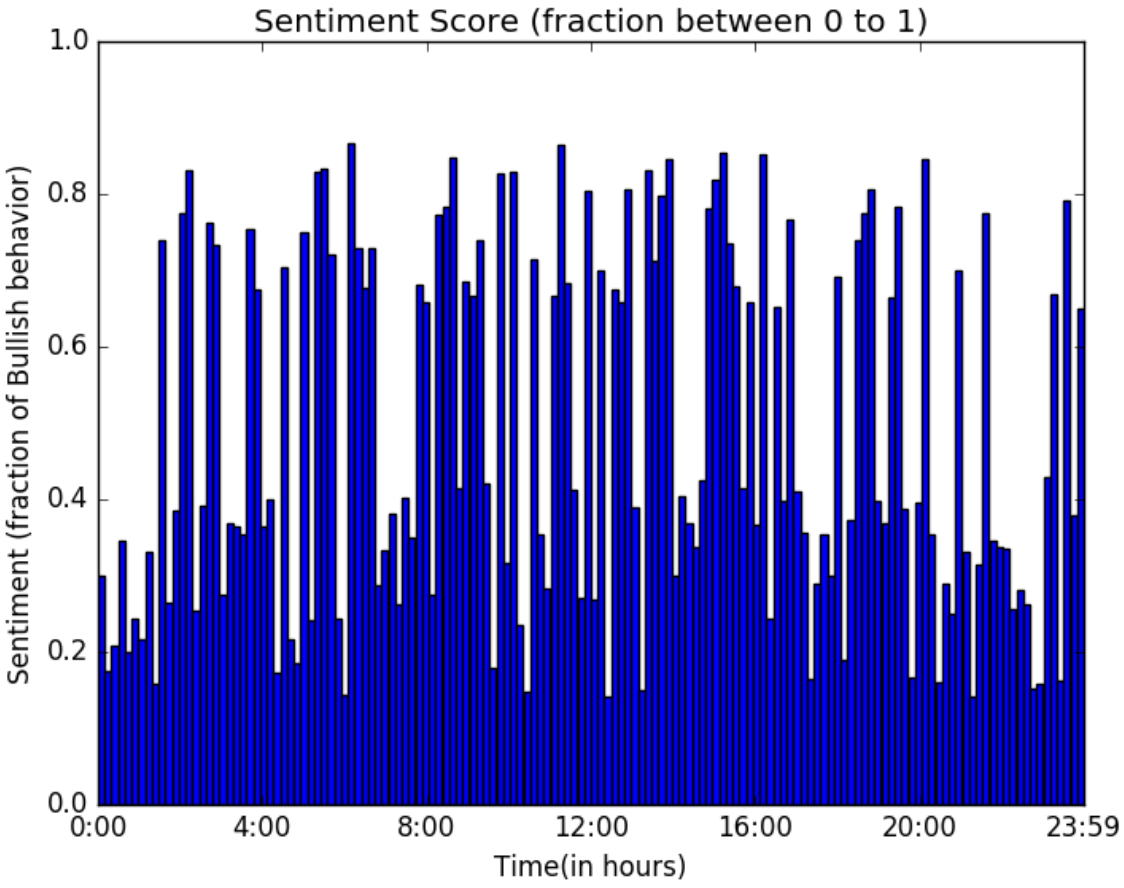

Figure 3.7: Sentiment Score Graph(Sample)

`

3.6.2 For Stock Price

For plotting the sentiments score of StockTwits messages, line graph is used to plot the data. To do this, Matplotlib (a python's matlab library) is used. The sentiment's graph y-axis varies between the highest price of AAPL and lowest price of AAPL while the x-axis varies from 9:00hrs to 16:00 hrs. This graph contains a line for average value of stock price for a day. This graph also contains the trend-line for stock market variation. A sample of stock market's graph is shown in figure 3.8.

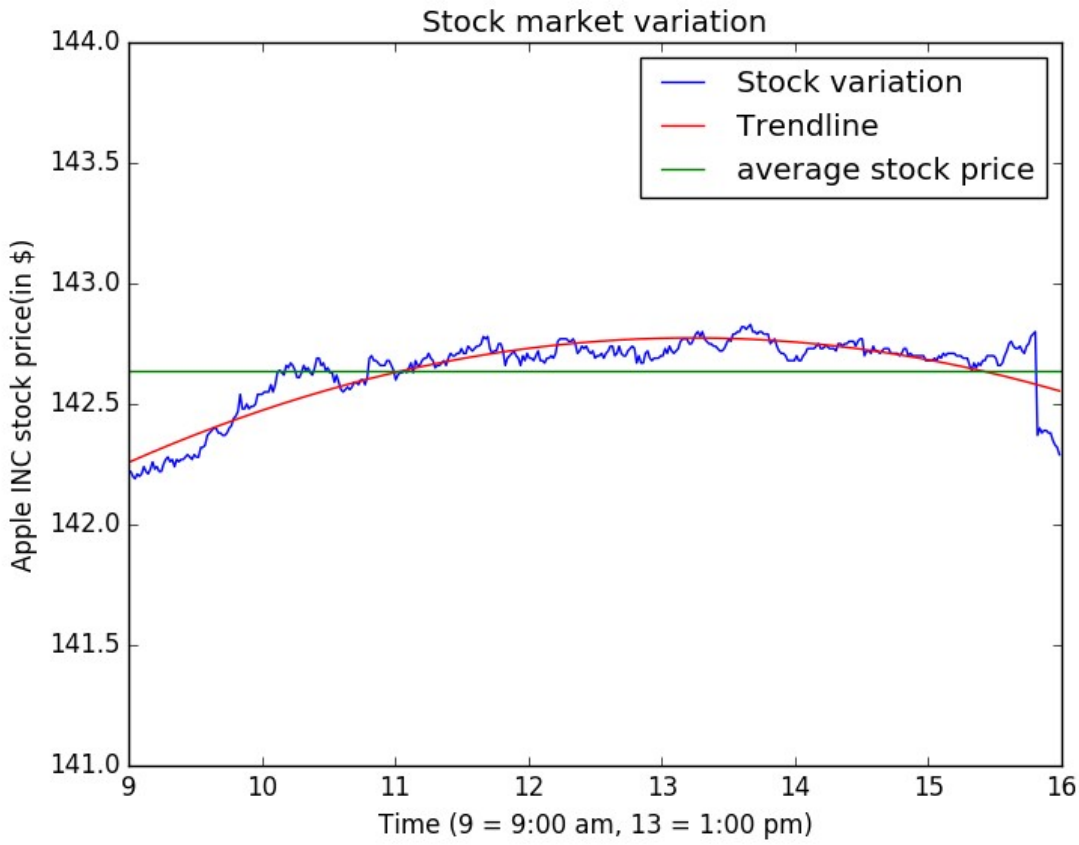

Figure 3.8: Stock Price Graph(Sample)

`

4. PERFORMANCE ANALYSIS

For training of the sentiment analysis tool, we have gathered around 4000 StockTwits messages through mining. We need to prepare the sentiment analysis engine and test on the same dataset. To do this, since we've rearranged our information set, we'll take out the initial 3,800 rearranged tweets, comprising of both positive and negative tweets, so that we can test against the last 200 to determine how precise we are.

This is known as directed machine learning, since we're demonstrating the machine information by letting it know about the sentiment of messages whether a particular sentiment is positive or negative in nature. This engine is stored in a .pickle file. Now we use the remaining data to test it.

Since Naive Bayes theoem is used for this procedure, the accuracy for sentiment analysis obtained by testing the data is 71.3%.

For the analysis of stock market with respect to StockTwits messages, 3 (3-4-2017 to 21-4-2017) weeks of data is obtained. The StockTwits messages data varies from day to day due to variation of user post in stockTwits for Apple Inc.

A sample of sentiment analysis done by the above discussed techniques is shown in Table 4.1:

Table 4.1 : Sample of StockTwits messages and their sentiments extracted by the above mentioned model.

Messages From StockTwits	**Sentiment Tag**	**Sentiment Score**
$AAPL bought 250 “today” 142 puts at $0.12 just Incase they pin 141. Not worried on the 142 calls; have next week.	Neg	0.384153066027
$AAPL it's very, odd. Was tracking DOW ok. Tanks with DOW, now being kept in very tight range. If they pin it & u can stomach risks, Mon +++	Neg	0.265798122642
$AAPL could be looking at a $142 pin today. I say this	Pos	0.80841504

`

Messages From StockTwits	**Sentiment Tag**	**Sentiment Score**
because they've excelled at keeping down after DOW recovered 65 points.		1485
$AAPL that crude settlement should be tanking the DOW. Only thing I can think of is Trump's tax plan release next week.	Pos	0.677949066486
$AAPL I actually think they test the waters about now, to make sure they can maintain the close (on option days).	Neg	0.228237978321
$AAPL we're taking off	Pos	0.900205398 61
$AAPL I'm going to hold those 142 puts to literally minutes before the close. It's just a couple of grand.	Neg	0.377395706184
$AAPL I wish the DOW would tank, drag apple way down, sell my puts, buy another 100 28th 141 calls below 1.80 & just wait until Monday.	Neg	0.254468923566
$AAPL who, aside from an algorithm, trades $142 a share stock at a half a penny between the spread of .01? .98 by . 99 and goes at .985. wow	Pos	0.719540856593
$AAPL BTW Grasshopper @adallica , I'm +$2,569 on the $141 calls. We'll see how Monday goes. I'll make sure 2 look u up since ur so concerned	Pos	0.799656253143

`

The graph for variation of stock market is shown in figure 4.1.:

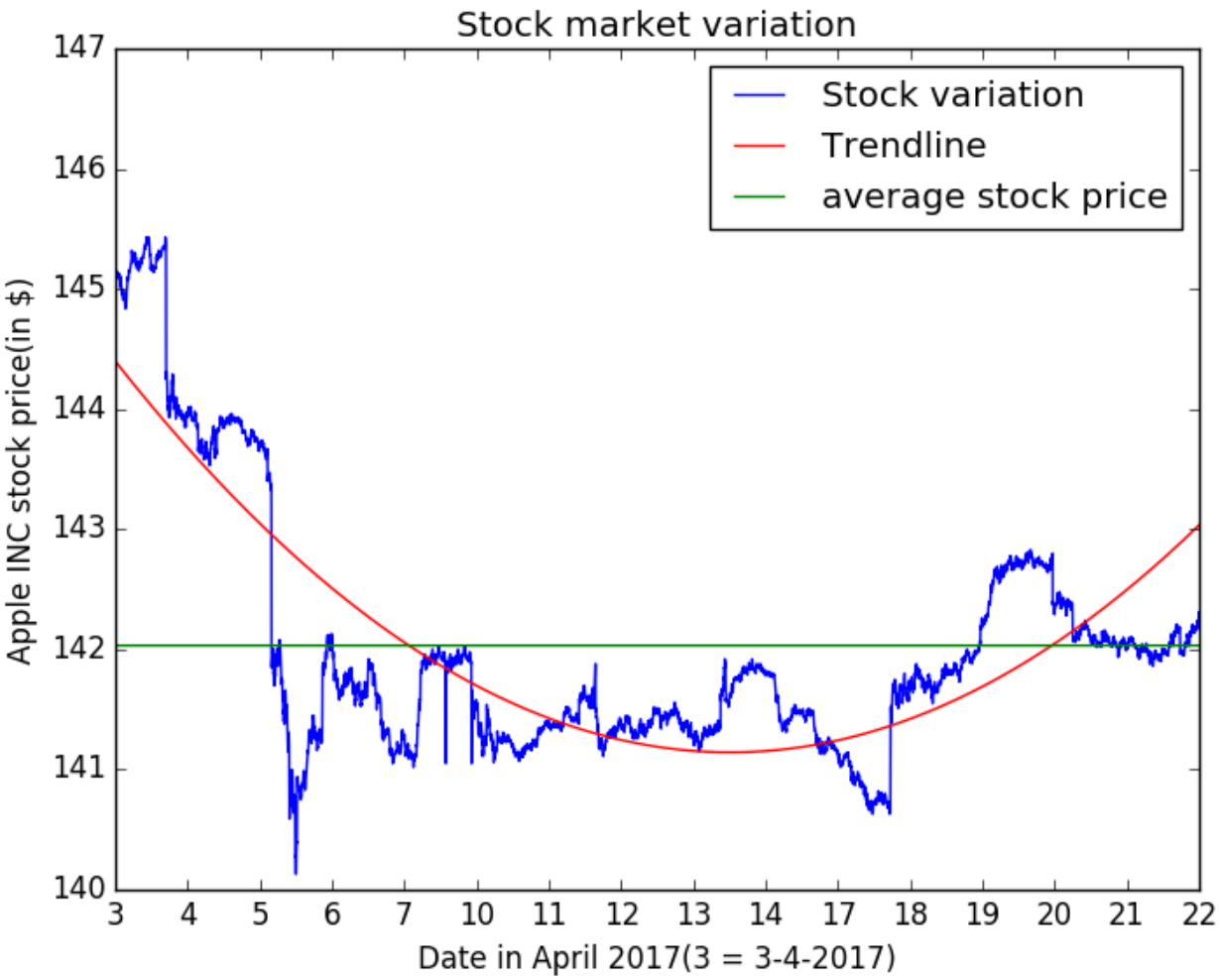

Figure 4.1: 3 weeks of stock market variation for Apple Inc

`

The graph for calculated 3 weeks of StockTwits messages' sentiments is shown in figure 4.2:

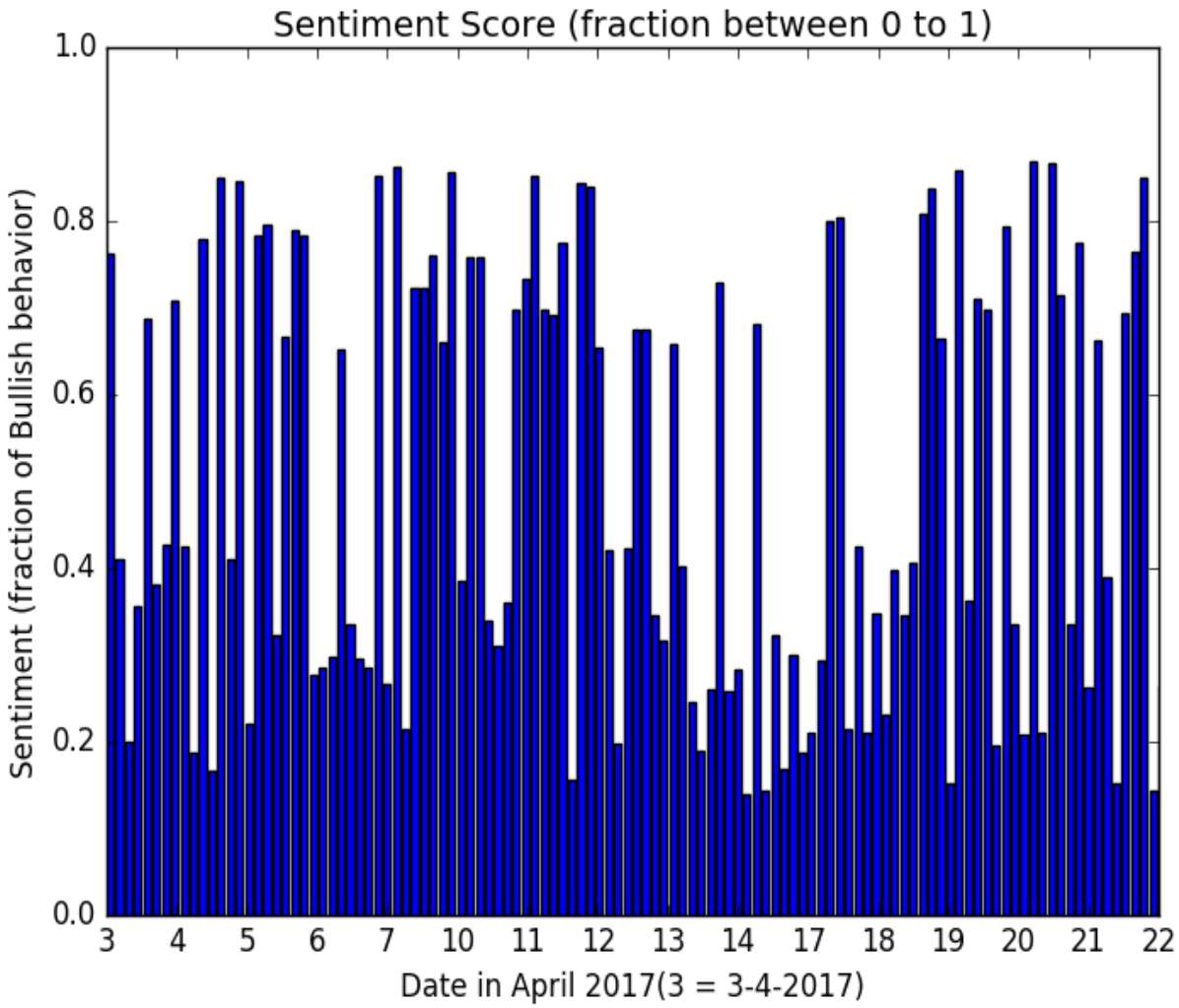

Figure 4.2: 3 weeks of StockTwits messages' Sentiment Graph regarding Apple INC

From the above two graphs, there is a sure result that the stock market of Apple Inc. follows the trend of post in the social media platform. It gave us result that the posts on social media has some affect on the fluctuation of stock price of Apple Inc. But we can say that the social media is mainly effective on predicting the fluctuation of stock market is only when there is negative behavior of posts and stock shows bearish behavior overall. It does not work well when there is a bullish behavior of stock market. The social media is not much effective on that behavior.

`

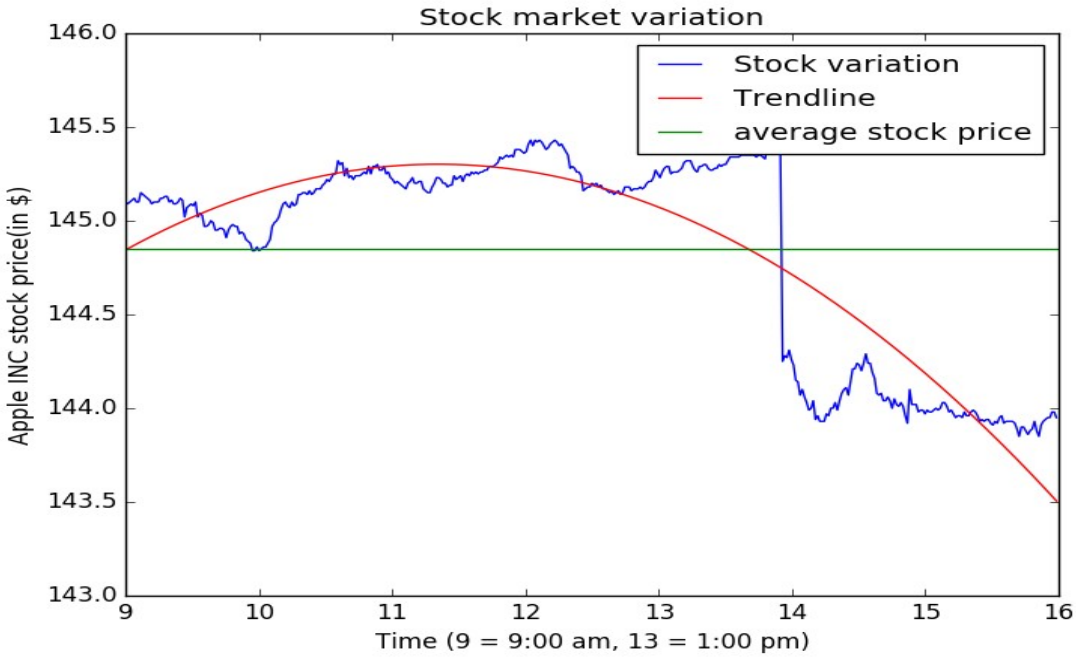

Figure 4.3: Stock market variation of Apple Inc on 3-04-2017

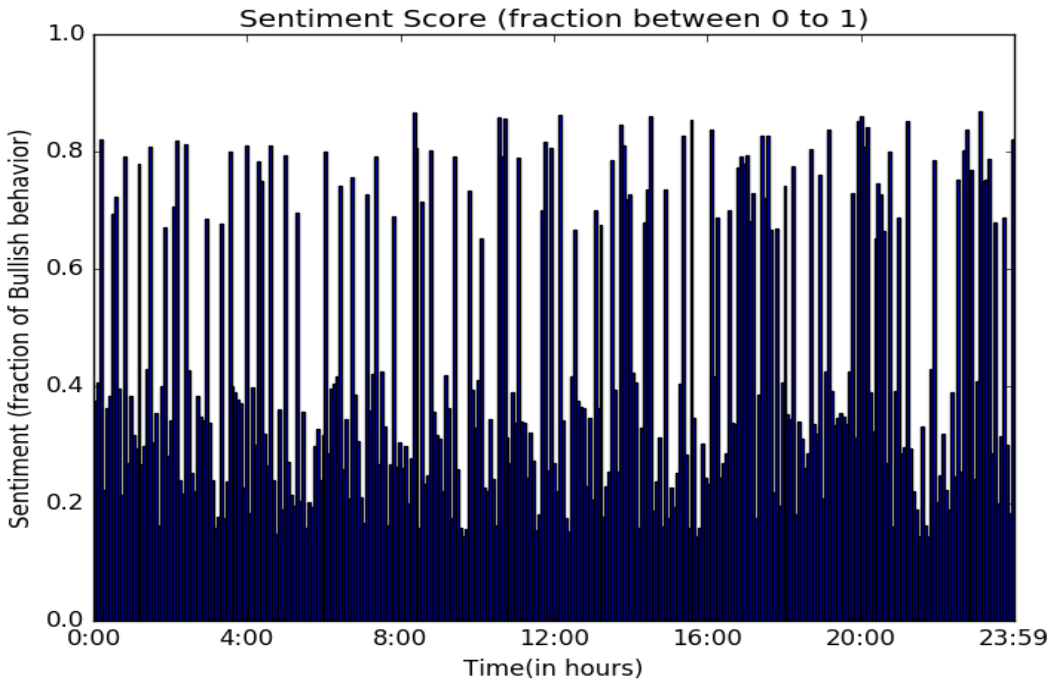

Figure 4.4: Sentiments of StockTwits on 3-04-2017

`

From figure 4.3 and 4.4, the stock price for Apple Inc shows bullish behavior at first but then there is drastic downfall in stock price which is due mainly due to the Internet's reaction is mostly negative in between those dates.

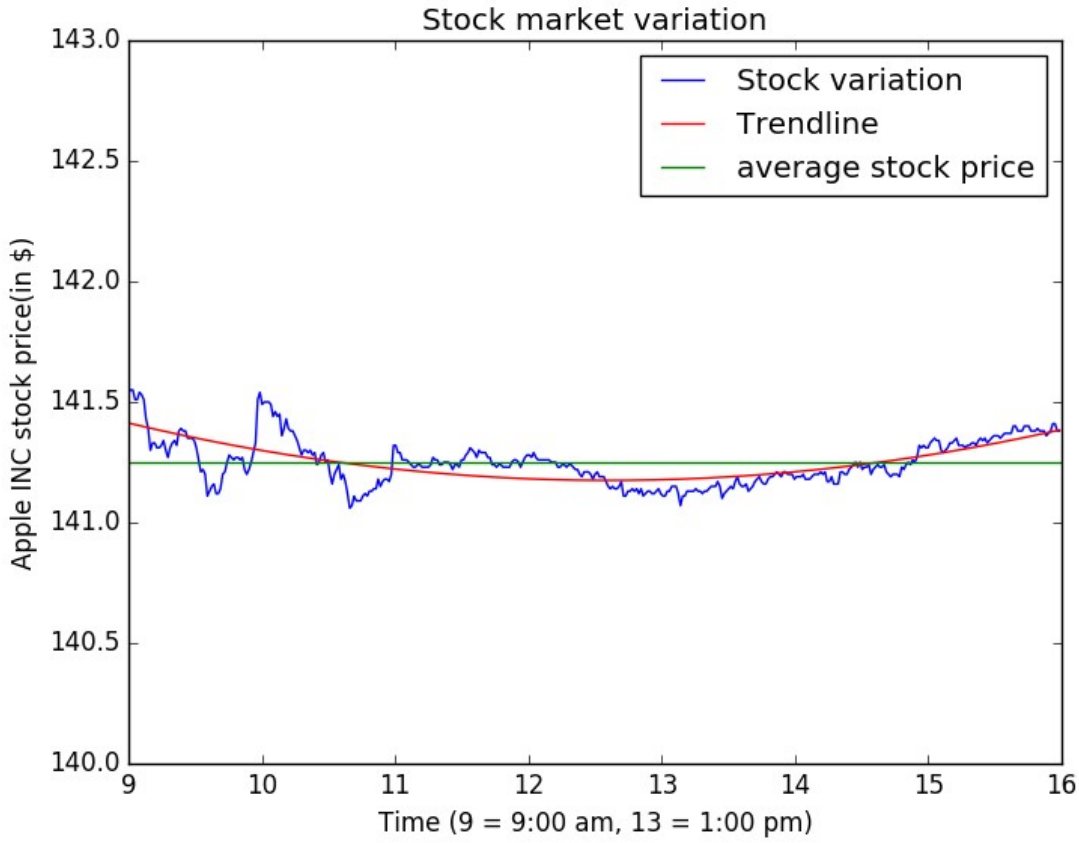

Figure 4.5: Stock market variation of Apple Inc on 10-04-2017

`

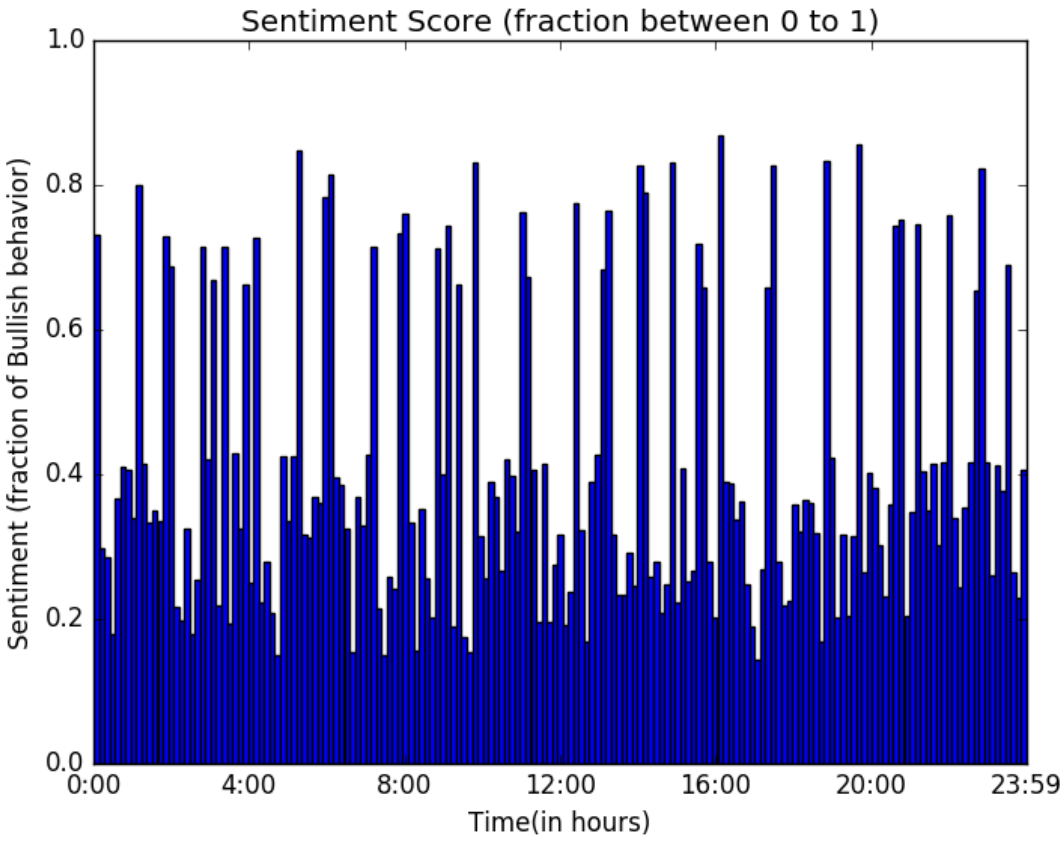

Figure 4.6: Sentiments of StockTwits on 10-04-2017

From figure 4.5 and 4.6, there is slightly change in stock price for Apple Inc is first decreasing the increasing in nature and this can be related to sentiment analysis of StockTwits messages behavior in between those dates is mixed for overall days.

From figure 4.1 and 4.2 ,in week 2, that is, between 10-4-2017 and 14-4-2017, the stock price for Apple Inc first increases then decreases and this can be related to positive behavior of sentiments at the start of the week, but as week progresses there is decrease in sentiments score of StockTwits' messages.

From figure 4.1 and 4.2, in week 3, that is, between 17-4-2017 and 22-04-2017, the stock price for Apple Inc first decreases the there is increments in it's stock price and this can be related overall positive sentiment score of StockTwits messages which may have result in the given manner of stock price change for this week.

`

Also there are some days in which the stock price does not follow the overall sentiments of StockTwits messages.

This analysis can be improved by increasing the sentiment analysis accuracy which can be done by using two or more algorithms by using a neural network to calculate sentiment score. Also this can be improved by using more amount of data and this lower accuracy in above performed analysis is mainly due under-fitting of data.

`

5. CONCLUSION

In the performed analysis we can relate the flow of stock price of Apple INC is somehow related the sentiments variation in StockTwits regarding Apple INC. Thus the posts in social network platform such StockTwits, Twitter, Facebook etc influences the changes in stock market. But the change in stock market is also depend upon a company's trading policies, its products success, launch of new product etc. This change is not solely dependent on social network post. Thus, along with social network posts and other factor that influence stock market can be used to predict the future change in stock market more accurately.

`

6. REFRENCES.

[1] Sang Chung, & Sandy Liu (2011, December). Predicting Stock Market Fluctuation from Twitter: An analysis of the predictive powers of real time social media

[2] Linho Zhang, (2013, April). Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation, Austin: Department of Computer Science, The University of Texas.

[3] Antoniou, G. C. (2012). Creation of a Twitter web-hint system that propose tweets based on the user preferences. Thessaloniki, Greece: Dept. of Electrical and Computer Engineering, Aristotle University of Thessaloniki .

[4] Nuno Oliviera, (2014). On the predictibility of Stock Market Brhavior using StockTwits Sentiment and Posting Volume

[5] Barbosa, L., & Feng, J. (2010, August). Robust Sentiment Detection on Twitter from Biased and Noisy Data. Coling 2010: Poster Volume, pp. 36-44.

[6] Ray Chen, & Marius Lazer. Sentiment Analysis of Twitter feeds for the prediction of stock market movement.

[7] SA Bogle, & WD Potter, (2015). SentAMaL – A Sentiment Analysis Machine Learning Stock Predictive Model. International Conference Artificial Intelligence, 2015. Georgia: Department of Computer Science, University of Georgia.

[8] Cavnar, W. B., & John, T. M. (1994). N-Gram-Based Text Categorization.
Environmental Research Institute of Michigan.

[9] Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., & Gandini, G. (2007). Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for

`

opinion mining. Andrea Sans , ed., Language Resources and Linguistic Theory: Typology, Second Language Acquisition, English Linguistics.

[10] Cianciullo, J. (n.d.). SocialMention. Retrieved Oktober 2013, from http://www.socialmention.com/

[11] Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., & Hughes, M. (2013). Sentiment Analysis of Political Tweets: Towards an Accurate Classifier. Proceedings of the Workshop on Language in Social Media (LASM 2013), (pp. 49-58). Atlanta, Georgia.

[12] Antoniou C. Georgios, (2013, November). Sentiment Analysis of Twitter Post. Thessaloniki, Greece. International Hellenic University

[13] Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. Coling 2010: Poster Volume, pp. 241-249.

[14] Dergiades, T. (2012). Do investors' sentiment dynamics affect stock returns? Evidence from the US economy. Economics Letters, 116(3), pp. 404-407.

[15] Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource.

[16] Ganter, B. B., & Wille, R. (1999). Formal concept analysis, mathematical foundation. Berlin: Springer Verlag.

[17] Gruber, T. R. (1993, June). A translation approach to portable ontology specifications. Knowledge Acquisition 5 (2), pp. 199-220.

[18] Hall, M., & Frank, E. (2001). A Simple Approach to Ordinal Classification. 12th European Conference on Machine Learning, (pp. 145-156).

`

[19] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 168–177).

[20] John G. Cleary, L. E. (1995). K*: An Instance-based Learner Using an Entropic Distance Measure. 12th International Conference on Machine Learning, (pp.
108-114).

[21] Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontologybased sentiment analysis of twitter posts. Elsevier(Expert Systems with Applications 40 (2013)), pp. 4065–4074.

[22] Anshul Mittal, & Arpit Goel, (2010). Stock Prediction using sentiment analysis. Stanford University.

[23] Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! Association for the Advancement of Artificial.

[24] Kumar, S., Morstatter, F., & Liu, H. (2013). Twitter Data Analytics. Springer.

[25] Liu, B. (2012). Sentiment Analysis and Opinion Mining. AAAI-2011, EACL-2012, and Sentiment Analysis Symposium.

[26] Liu, Y., Huang, X., An, A., & Yu, X. (2007). ARSA: A Sentiment-Aware Model for Predicting Sales (Vol. SIGIR’07). Amsterdam.

[27] Miller, G. A. (1995). WordNet: A lexical database for English. Communications of the ACM 38(11), pp. 39–41.

`

[28] Naes, T., Isaksson, T., Fearn, T., & Davies, T. (2002). A User Friendly Guide to Multivariate Calibration and Classification. NIR Publications.

[29] Nebhi, K. (2012). Ontology-Based Information Extraction from Twitter. Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data (pp. 17-22). Mumbai: COLING.

[30] Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity. ACL ’04, (pp. 271–278).

[31] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning. Proceedings of EMNLP, (pp. pp. 79–86).

[32] Park, S., Ko, M., Kim, J., Liu, Y., & Song, J. (2011). The Politics of Comments:

[33] Predicting Political Orientation. ACM 978-1-4503-0556-3/11/03. Hangzhou.
Stanford University. (2013). The Stanford Natural Language Processing Group. Retrieved from http://nlp.stanford.edu/software/corenlp.shtml

[34] Stone, P. J., Dunphry, D., Smith, M., & Ogilvie, D. (1966). The General Inquirer: A Computer Approach to Content Analysis. Cambridge: MIT Press.

[35] Thessaloniki: Dept. of Informatics, Aristotle University of Thessaloniki.

[36] Tweet Sentiment Visualization. (n.d.). Retrieved Oktober 2013, from Sentiment Viz: http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

[37] Waikato, D. O. (2013, Oktober). Weka 3: Data Mining Software in Java. Retrieved from http://www.cs.waikato.ac.nz/~ml/weka/

`

[38] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining, Practical Machine Learning Tools and Techniques (3rd ed.). U.S.A.: Morgan Kaufmann Publishers.

[39] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., . . . Steinberg, D. (2007). Top 10 algorithms in data mining. Springer-Verlag.

[40] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining Lexiconbased and Learning-based Methods for Twitter. HP Laboratories(HPL-2011-89).

[41] Zhang, L., Xu, W., & Li, S. (2012). Aspect identification and sentiment analysis based on NLP. Network Infrastructure and Digital Content (IC-NIDC), 2012 3rd IEEE International Conference on, (pp. 660 - 664). Beijing.