Predicting stock market movements using natural language processing and regression as a tool

Thesis submitted in partial fulfilment of the requirement for the award of Degree of

Bachelor of Technology

in

Electronics and Communication Engineering

Submitted by

SUNANDA SAHA
(Roll No. **15/EC/44**)

Under the supervision of

AURPAN MAJUMDER

(Assistant Professor)

Department of Electronics and Communication Engineering

National Institute of Technology, Durgapur

May 2019

NATIONAL INSTITUTE OF TECHNOLOGY DURGAPUR

CERTIFICATE

It is certified that the work contained in the thesis entitled **"Predicting stock market movements using natural language processing and regression as a tool"** has been carried out by **Sunanda Saha (15/EC/44)** under the guidance of **Mr. Aurpan Majumder,** the data reported herein is original and that this work has not been submitted elsewhere for any other Degree or Diploma.

Sunanda Saha
15/EC/44
Place: **Department of Electronics and Communication Engineering, NIT Durgapur**
Date: **10/05/2019**

This is to certify that the above declaration is true.

Mr. Aurpan Majumder
Assistant Professor
Department of Electronics and Communication Engineering, NIT Durgapur
Date: **10/05/2019**

Acknowledgement

I have no words to express my gratitude and thanks to Mr. Aurpan Majumder, Assistant Professor, Electronics and Communication Engineering Department, NIT Durgapur, for his precious guidance and effectual care which happens to be the psyche of this thesis report. I consider them as my great advisers and will continue to seek their guidance in future accomplishments.

The thesis report couldn't be furnished without experience and versatility so I would like to express my heartily thanks to all the respected professors of Electronics and Communication Engineering Department for their valuable technical and moral suggestion and also constant encouragement, without which this thesis report would not come in to existence.

I am grateful to express my gratitude and appreciation to my friends, batch mates, all those who has provided many helpful suggestions and also encouraged me from time to time in completion of this project.

Finally, I would like to thank my parents and members of my family for their motivation and encouragement at all times.

Date: **10/05/2019**
Sunanda Saha
Roll No: **15/EC/44**
Department of Electronics and communication Engineering
National Institute of Technology Durgapur
West Bengal-713209, India

Contents

Chapter 1

Introduction

Social networking services have become more than just a tool to connect with friends. There are millions of users on these websites and these users generate vast amounts of data. Twitter is an example of such a social network platform, where the users are connected with each other in a unidirectional manner. That is a link does not mean two users are friends with each other. Users which follow a particular user are called followers of that user. Thus, it is not necessary that a user A being followed by B should also follow back B. In the last few years, the Twitter has transformed from its original intended purpose of a simple, personal, a micro-blogging site to mega content generator. Twitter has approximately closed to 300 million active users and about 500 million of posts, which are also called tweets, are generated by users, per day.

Over the time, Twitter has become a fundamental source of information for news. A post generated by a user and its subsequent viral propagation in the network, even has attracted well established news sources. As a one step forward, researchers have tried to analyse if the tweets contain predictive power. This has resulted in various works in studies in different fields, such as the study of the spread of epidemics, the prediction of electoral results or of football matches results. In particular in the financial domain, a lot of research has been done to find if there exists a correlation between the tweets and the trend of the stock market. In other words, the researchers tried to investigate if tweets can affect (or predict) the financial market.

In this thesis, we present our study about understanding and findings of relation between tweets and stocks. In twitter with respect to financial tweets, users can embed cashtags, which are stock symbols embedded immediately after the $ sign. Using cashtags, users can search about the tweets related to a particular stock.

1.1 Problem Statement

In the past a lot of research has been done to propose a function, which takes as input all the tweets for a particular stock or indexes, analyse them and predict the stock or index price. In this work, we take an alternative approach: using the stock price and tweet information, we investigate following questions.

- Is there any relation between the sentiment of the tweets and stock prices?
- If there is a relation what is the structure of the graph that describes the relationship?

We take the help of sentiment analysis which lies within the domain of Natural language processing.

1.2 Contribution

In this thesis, we analyze the vast volume of data and present our quantitative and qualitative correlation results about the relation between stocks and tweets. For doing so we have built a classifier using logistic regression. Logistic Regression is a good baseline model to use for several reasons: (1) They're easy to interpret, (2) linear models tend to perform well on sparse datasets like this one, and (3) they learn very fast compared to other algorithms. To keep things simple I'm only going to worry about the hyper parameter C, which adjusts the regularization. The structure of the graph represents the relationship between the actual and the predicted values of stock.

Alongside, we also diversified our work over ten top grossing companies in the world most of which were listed in Forbes 500 and also classified them over a span of sixteen years (2001 – 2015).

The rest of the report is organized as follows. Section 2 describes some terminologies with respect to Twitter. In section 3 we present related literature with respect to our work. We then discuss about the algorithms used in section 4. The methodology and approach used is detailed in Section 5. This section also presents our results related to the dataset we analysed. We conclude with several future directions in section 6.

Chapter 2

Preface

In this chapter, we give a basic background with respect to Twitter and in particular tweets related with finance.

2.1 Twitter

As with many stories of success, even more in the case of social networking services, Twitter's history is very complex and controversial. Since its launch in 2006, initially called "twttr", allowed its users to write a short message of 140 characters. Doing this operation of status updating, is widely called "to tweet".

The relations in Twitter are unidirectional that is an edge from user A to B, not necessarily means vice versa. Thus, there is no concept of bidirec-tional friendship as present in social networks like Facebook. An user A can subscribe to another user B's feed: the operation is termed as "following". Thus, A is called a follower and B is called followee, in Twitter terminologies.

A user can repost someone else's tweet on its personal feed, which is termed as "retweeting" and interact with other users, replying to a tweet or tagging another user in a tweet with the operator "@". Another key operation of Twitter, present since the beginning, is the tagging a tweet using keywords preceded by the symbol "#". Keywords of such kind are called hashtags and they're the base of the research on Twitter by topics. One of the key features of Twitter is the trending topics. This feature is geo specific and exploits whom a user is following.

These features, conjunct to its typical, as said before, directed nature, unlike, for instance Facebook, which is bidirectional, contributed to ascend Twitter to the role of most popular social networking site for news, politicians, celebrities and influent bloggers.

2.2 Twitter for Finance

In July 2012, Twitter introduced the possibility for the users of making a search using the ticker symbol of a stock, preceded by the symbol "$". Such a symbol, or operator, for instance $AAPL, is also called cashtag. This feature was introduced for the first time in 2008 in Stocktwits, which is equivalent of Twitter however, only related to financial domain.

The Twitter users with respect to financial domain can be categorised in the following three categories:

1. Financial news channels: These users drive information linking in their tweets articles from their site.
2. Trading bloggers: These sets of users express their opinion on the stocks.
3. Investors: They are users that simply follow others users for advice.

Intuitively, users of third categories are generally followers of first and second types. One of the aims of this thesis is to analyse the network structure and to cross validate our intuition.

Chapter 3

Related Works

In this section, we present various related works with respect to our work in section 3.3. As mentioned, our work is mainly related to analysis of financial tweets and correlating the tweets with the fluctuations in the stock market. However, this work is very much related to prediction of stock market using sentiment analysis. Thus, we present works done in the past, where authors have proposed various approaches for prediction of various entities in general in section 3.1 and then in particular with respect to financial domain in section 3.2.

3.1 Prediction

Researchers have always been fascinated by large amount of data on the web and exploiting and analysing it to predict various entities such as for football games, consumer behaviour, music sales, elections, epidemics, movies sales.

Here we discuss the works with respect to Twitter data. Authors proposed a predictive model tested using cross validation, to predict football outcomes. They used both Twitter and historical data for analysis and reported a 75% accuracy of the outcomes. To predict the movie box office results, a model based on the tweets creation rate is proposed in. The approach uses sentiment analysis and claim to outperform market-based pre-dictions. Researchers try to predict elections using Twitter datasets by performing a sentiment analysis of the tweets. The method exploits the mentioning of party/politician and political coalitions. Apart from predictions, Twitter data is used for tracking an epidemic. Authors described a tracking system for tracking the prevalence of Influenza-like Illness (ILI) in several regions of the United Kingdom.

3.2 Prediction in Financial data

In this section, we present literatures, which are related to prediction of stock market using Twitter data analysis.

Prediction of the stock market using web data such as blogs and other web social platforms specially Twitter has attracted a lot of researchers in the past. Using SVM based technique, authors predicts the stock market with 87% accuracy. However, the data on blogs is small compared to Twitter. This motivated later studies to use Twitter data extensively for stock market prediction.

In the initial set of studies in Twitter dataset, which is done over a six month period, authors identify the tweets into six sets of moods and then use it for prediction of change in values in the Dow Jones Industrial Average (DJIA). Later in a different work, researchers argued that it is important to find experts in the dataset to predict the values. However, in this approach a lot of users who are not followers and make their opinions lost their voice. By using sentiment analysis, researchers proposed a model to predict the stock market. Their aim was to investigate if there is any correlation between Twitter and the stock market by studying sentiment, message volume, price movement and stock volume as well as the effect that a Twitter user's reputation may have on sentiment and the stock market. In the later part of the researches, researchers exploit sentiment analysis clubbing with machine learning techniques in on Twitter corpus. They all find out a strong correlation among tweets and stock markets.

3.3 Analysis of financial data

In this section, we present works which have also analysed the correlation between financial markets and the Twitter activity. The authors study the correlation between tweets and stock market events such as changes in price and traded volume of stocks. There are many differences between their work and this thesis.

- Furthermore, their research is focused on the interaction graph that they build based on the tweets, while we investigate also on the relationship between users and on the relationship between stocks.
- Another difference is that we also performed the sentiment analysis of the tweets and its effect on the stock market, which is missing in their work.

Another work, where researchers analysed the set of tweets to understand the mood of tweets and to relate it to the stock price is done. However, the main emphasis of their work is about sentiment analysis and its effect on the stock market in general. Specifically, they evaluated the tweets from August 1, 2008 to December 20, 2008 and considering 8 events. In our case the time line is much longer. However, the main difference is that in their investigation they investigated if socio-economic phenomena (such as peaks in stock market or oil price, Presidential Election, Thanksgiving day) affect the public mood, obtained from Twitter, while we're looking in the opposite direction: if Twitter mood is useful to predict economic phenomena.

- In another similar work, researchers analyzed tweets to find correlations among tweets and stock market in various sectors. However, their dataset is less than of three months, whereas our dataset spans over ten years. Also, as mentioned before, we are targeting the tweets containing cashtags. Also, compared to their work, where they proposed a model, we have also performed sentiment analysis to find relation between mood of the tweets with the stock fluctuation.

Chapter 4

Algorithms Used

4.1 Natural Language Processing

Natural Language Processing (NLP) is the best way to understand the language used and uncover the sentiment behind it. People often consider sentiment (in terms of positive or negative) as the most significant value of the opinions users express via social media. However, in reality emotions provide a richer set of information that address consumer choices and, in many cases, even determines their decisions. Because of this, Natural Language Processing for sentiment analysis focused on emotions is extremely useful.

4.1.1 Sentiment Analysis

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just scratching the surface and missing out on those high value insights that are waiting to be discovered.
With the recent advances in deep learning, the ability of algorithms to analyse text has improved considerably. Creative use of advanced artificial intelligence techniques can be an effective tool for doing in-depth research. Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral.

4.1.2 Usefulness of Sentiment Analysis

Sentiment analysis solves a number of genuine business problems. It helps to predict customer behaviour for a particular product. It can help to test the adaptability of a product. It automates the task of customer preference reports. It can easily automate the process of determining how well did a product run by analyzing the sentiments behind the user's reviews from a number of platforms.

4.2 Logistic Regression

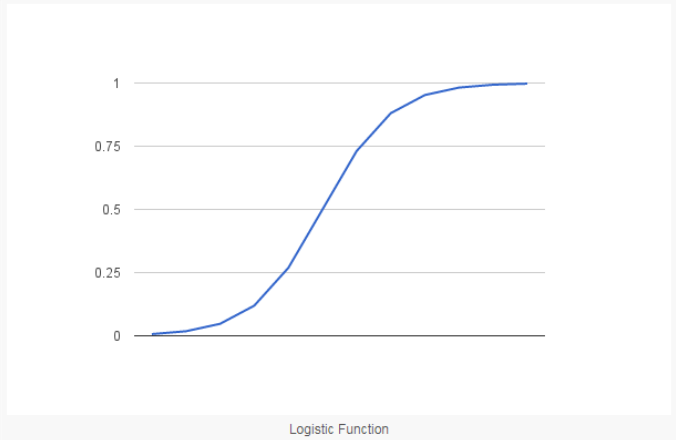
Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values).

4.2.1 Logistic Function

Logistic regression is named for the function used at the core of the method, the logistic function.
The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

Where e is the base of the natural logarithm and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function. Now that we know what the logistic function is, let's see how it is used in logistic regression.



4.2.2 Representation used for logistic regression

Logistic regression uses an equation as the representation, very much like other regressions. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modelled is a binary value (0 or 1) rather than a numeric value. Below is an example logistic regression equation:

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data. The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b's).

Chapter 5

Evaluation

In this chapter, we first describe the dataset we used for our analysis then our approach and later, we describe various quantitative results of our investigation on the dataset.

5.1 Dataset

5.1.1 Kaggle

This is a hub for thousands of datasets specifically used for machine learning purposes. The dataset contains tweets with respect to 10 companies starting from 2001 to 2015 and is of size 946 MB. It contains approximately one million tweets which were tweeted by approximately 0.5 million unique users. Refer to Table 5.1 for summarization of the Twitter dataset being analysed.

5.1.2 Yahoo Finance

For particular months, we analyzed tweets with respect to nineteen companies; we also downloaded corresponding stock values from the historical archive of Yahoo Finance. Yahoo Finance stores historical financial data for all the companies quoted at the New York Stock Exchange (NYSE) and National Association of Securities Dealers Automated Quotation (NASDAQ). The site provides an API and a web interface to download the data from its database in a .csv format. The fields that we got, for every day, are Date, Open, High, Low, Close, Volume and Adjusted Close, that means the close value adjusted considering the eventual dividends and splits. We simplified it and then used the simplified dataset.

Company Number	10
Size	946 MB
Tweets Number	1054.164
Users Number	516.371
Date	2001 to 2015

Table 5.1

5.2 Approach

5.2.1 Building the classifier model

For this analysis we'll be using a dataset of 50,000 tweets. The data is split evenly with 25k tweets intended for training and 25k for testing our classifier. Moreover, each set has 12.5k positive and 12.5k negative tweets. Deliberately we chose a dataset which had equal positives and equal negatives to help make our process easier. The raw text is pretty messy for these tweets so before we can do any analytics we need to clean the tweets up and pre-process them; which include removing hyperlinks, tagged twitter users and other special characters from the tweets.

Next we move on to vectorization of the tweets. In order for this data to make sense to our machine learning algorithm we'll need to convert each review to a numeric representation, which we call vectorization. The simplest form of this is to create one very large matrix with one column for every unique word in your corpus (where the corpus is all 50k tweets in our case). Then we transform each review into one row containing 0s and 1s, where 1 means that the word in the corpus corresponding to that column appears in that review. That being said, each row of the matrix will be very sparse (mostly zeros). This process is also known as one hot encoding.

Now that we've transformed our dataset into a format suitable for modelling we can start building a classifier. Logistic Regression is a good baseline model for us to use for several reasons: (1) They're easy to interpret, (2) linear models tend to perform well on sparse datasets like this one, and (3) they learn very fast compared to other algorithms. To keep things simple we're only going to worry about the hyper parameter C, which adjusts the regularization.

After finding the optimal value for C, we train a model using the entire training set and evaluate our accuracy on the 25k test tweets.

ALGORITHM 1: Building classifier model

1. Clean and pre-process the tweets – remove hyperlinks, tags and special characters
2. Clean tweets are one hot encoded for readability of the our machine learning algorithm
3. n← total number of training tweets
4. for i← 1 to n/2 train model for 1 (positive tweets)
5. for i ←n/2 + 1 train model for 0 (negative tweets)

6. c←regularization constant varied from 0 to 1
7. Check accuracy (Highest for c = 0.05)
8. m← total number of testing tweets
9. test model for m tweets with c=0.05
10. Check accuracy (0.8876)

5.2.2 Applying the classifier model

Once our final model is ready we start our process of sentiment analysis. We arrange the dataset by tweets corresponding to each company and the year it was tweeted. The tweets are then extracted from our dataset, vectorized like the previous train and test tweets and finally the final classifier is applied to the tweets. The result returned by the model is a list of ones and zeros; ones indicating positive tweets and zeros indicating negative tweets. We sum the list returned; that is actually keeping a count of the positive tweets; and store it in a ten cross ten 2D matrixes, each cell representing a particular company and year. Next we add another column to the actual stock price dataset called 'predicted' and append the values of the previous matrix one by one. Our final step is plotting the actual stock value across the predicted value in the same plot to see the difference.
A very important thing to notice is that the actual stock value is in billions whereas that predicted by the tweets are in mere thousands. Thus a huge difference of scale exists.

ALGORITHM 2: Applying the classifier for sentiment analysis

1. Arrange dataset by grouping tweets according to the company in increasing order of the year
2. Define a function which takes in the dataset and performs the following
3. Create a ten cross sixteen 2-d integer array
4. r←each row of the dataset
   x←first company name
   y←first year
5. Taking in every row r of the data set one by one
6. If r.company_name equals x
7. If r.year equals y
8. Append the tweet to a list
9. End of inner if loop
10. Clean the list of tweets, vectorize them and apply it to the final model
11. The result gotten back is a list of ones and zeros
12. Add the numbers and divide it by its total size
13. y←r.year
14. End of outer if loop
15. x←r.company_name
16. Return 2-d array
17. End of function

Using the result for plotting our graphs

1. Append a new column ‘predicted’ to the dataset containing stock value
2. Append the rows of the new column with the values of the 2-d array
3. Define a function which takes in the company_name and performs the following
4. r←each row of the dataset
5. Taking in every row r of the data set one by one
6. If r.company_name equals the function argument
7. Plot a graph between r.actual and r.year
8. Plot another graph on the same plot between r.predicted and r.year
9. End of function

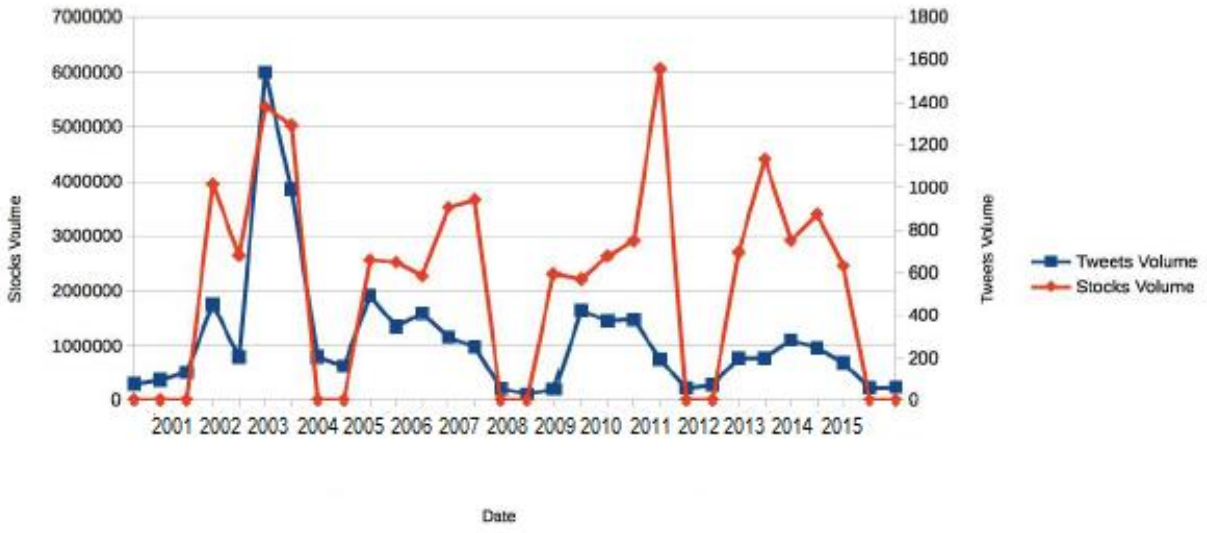
5.3 Results

We provide results related to top ten most tweeted stocks, which ranges from technology companies (Apple, Microsoft) to banking (Hsbc, Goldman Sachs) to e-commerce (Amazon) to entertainment portals (Netflix). We included a few more companies which made their place in the Forbes 500.

5.3.1 Quantitative analysis

One of the ideas of this thesis is to compare the daily volume of stocks traded with the volume of the tweets that has been tweeted in the same days, to see if there is some kind of correlation between the two values. We obtain information about the stock price from Yahoo Finance and the volume of the total stocks traded per year was compared to per year tweet volume for each particular stock.
For all the ten companies, we perform our analysis for the year 2005 to 2014.

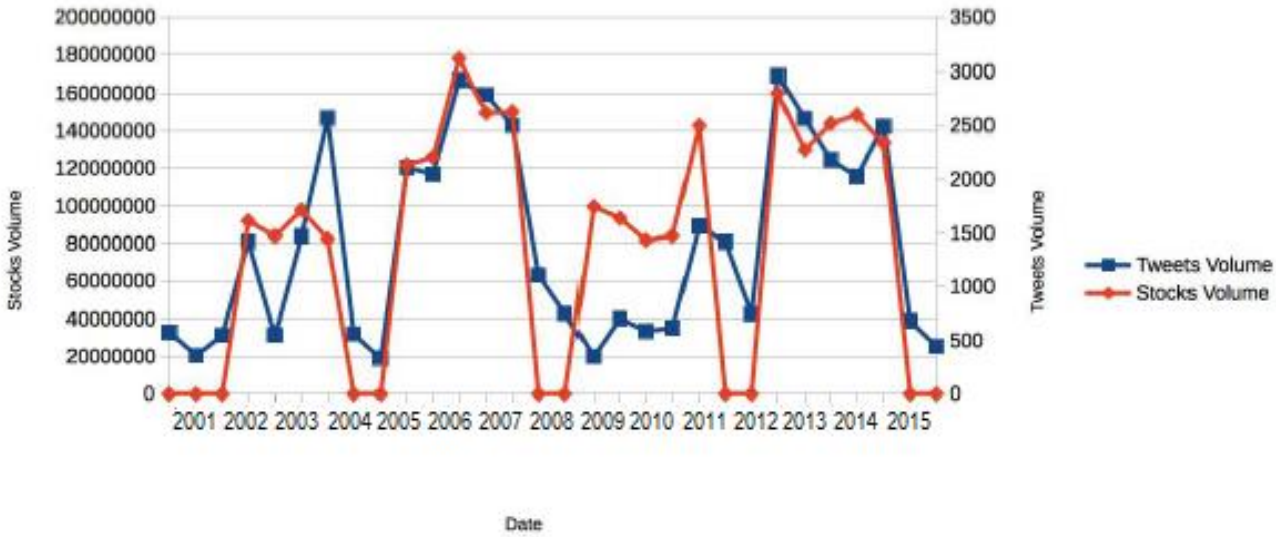
Alphabet:



As we can see that here the two curves vary by a great amount. The only place they sort of coincide is between 2002 and 2004. After that they seem to converge

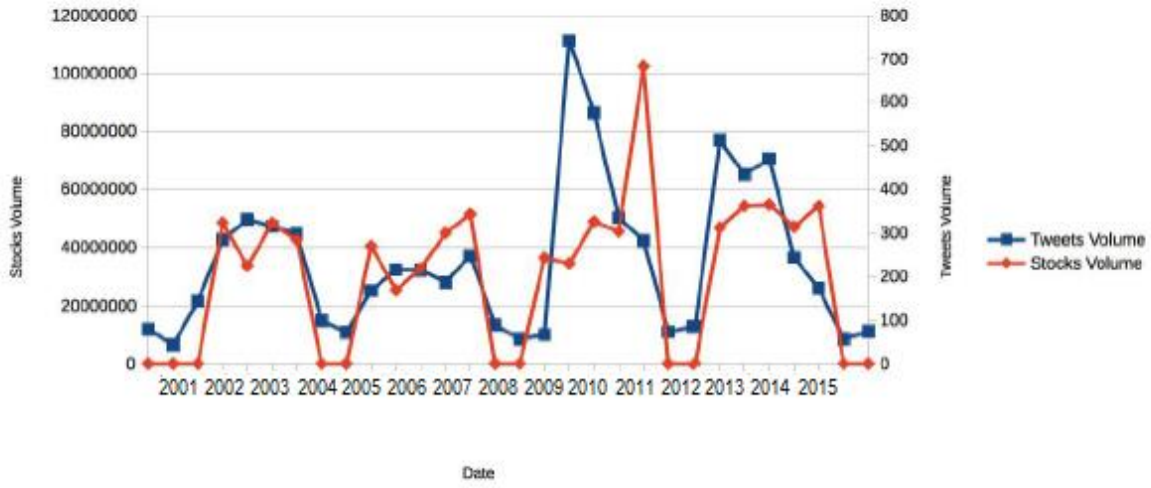
at a few years like 2008, 2012 and 2015 in between the rest of the years. We get only 33% efficiency for Alphabet with this model with regularization coefficient of 0.5. Other values of regularization coefficient could have produced better results.

Apple:



As seen here we get a slightly better model for Apple compared to Alphabet. Here too we have got great deviations especially between 2008 to 2012 but also we can see that two graphs nearly coincide from 2005 to 2008 and again from 2012 to 2015. We get an efficiency of 52% for Apple with this model with regularization coefficient of 0.5. Other values of regularization coefficient could have produced better results.

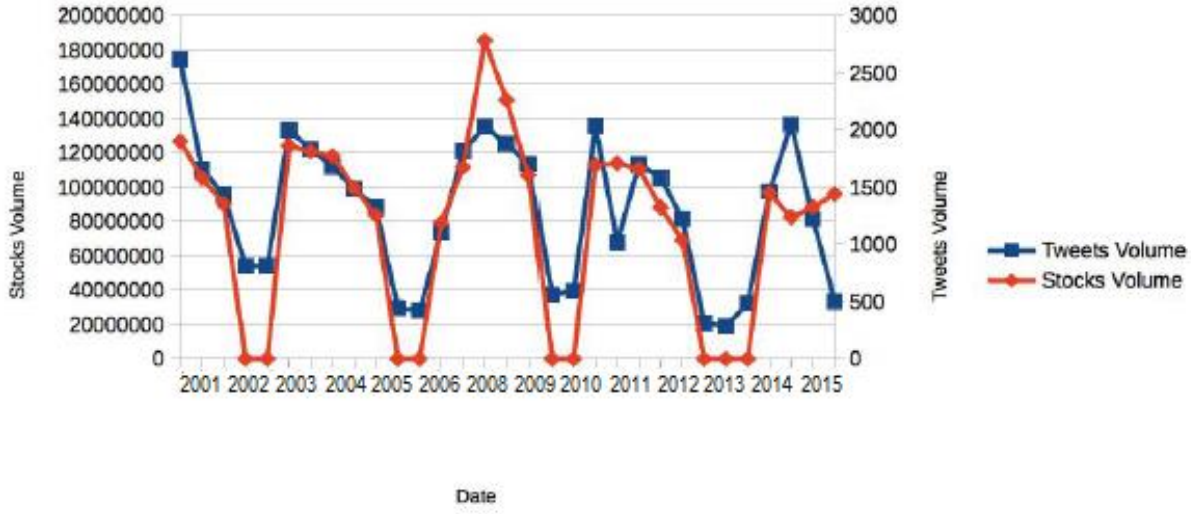
Amazon:



We get an even better result for Amazon compared to Alphabet and apple. Initial half of the graph has both the lines nearly coinciding. We see certain deviations in the third quarter and even slighter deviations in the final quarter. Amazon gets an

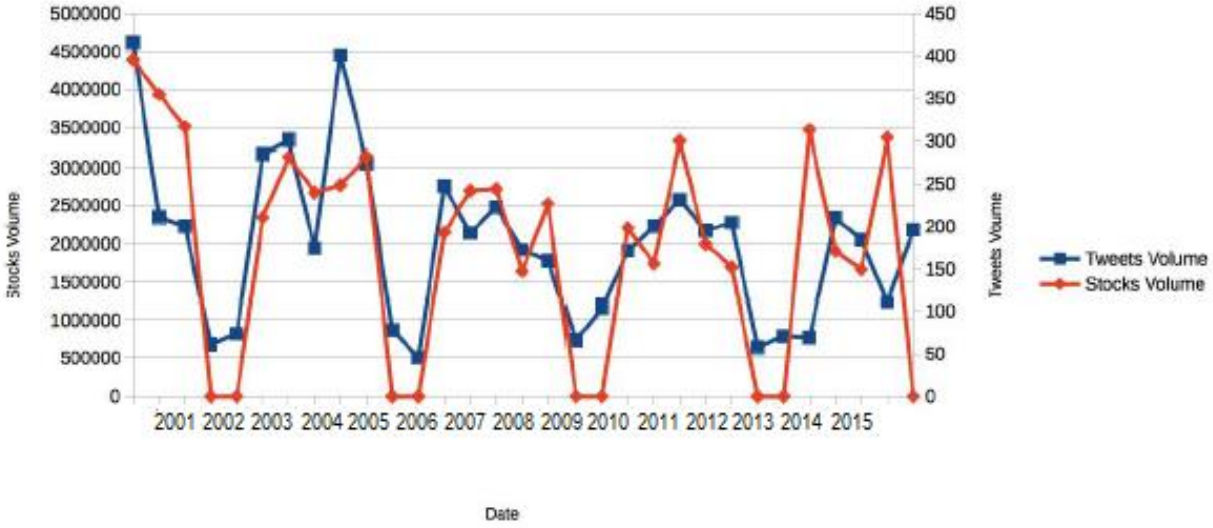
efficiency of 68% with this model with regularization coefficient of 0.5. Other values of regularization coefficient could have produced better results.

Goldman Sachs:



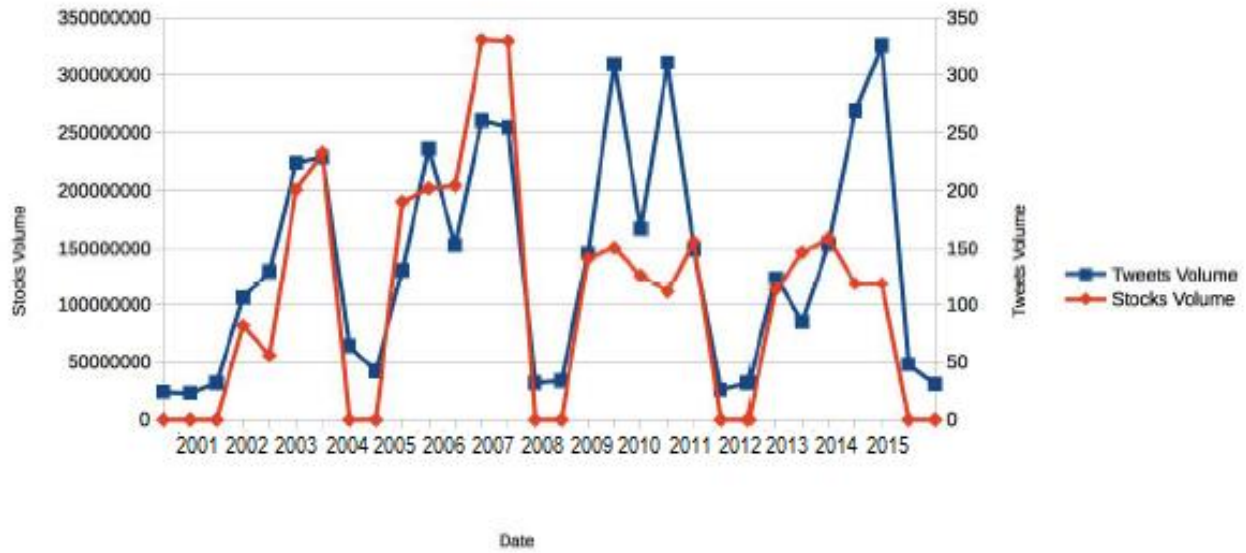
We can see that this is probably a good model for Goldman Sachs as most of it coincides with a few deviations here and there for a year or two. The model got high efficiency of 87% for Goldman Sachs with regularization coefficient of 0.5. Other values of regularization coefficient could have produced even better results.

Google:



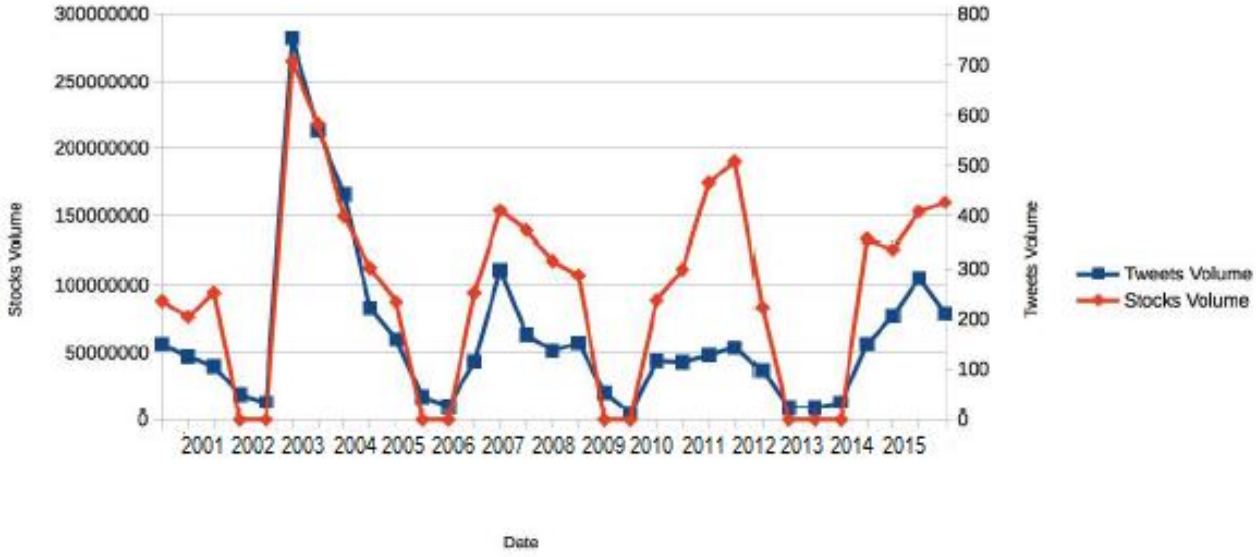
Our model gives a pretty good efficiency of 79% for Google. It's clearly visible that throughout, the two lines have lied pretty close to each other but never really coinciding at any place but still it fetches a pretty good prediction. Other values of regularization coefficient could have produced better results.

Hsbc:



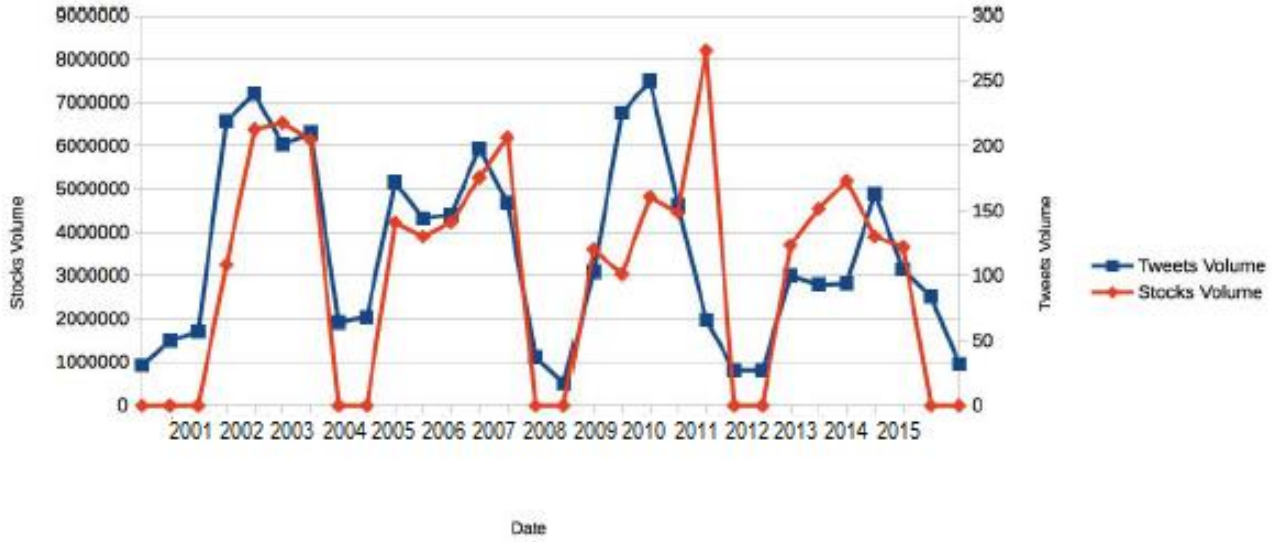
The first and second quarter prediction is more or less to the point but the third and the fourth quarter pull down the efficiency of the model for Hsbc. Third quarter is predicted totally wrong. Efficiency for Hsbc given by the model is 63% with regularization coefficient of 0.5. Other values of regularization coefficient could have produced better results.

J.P. Morgan Chase:



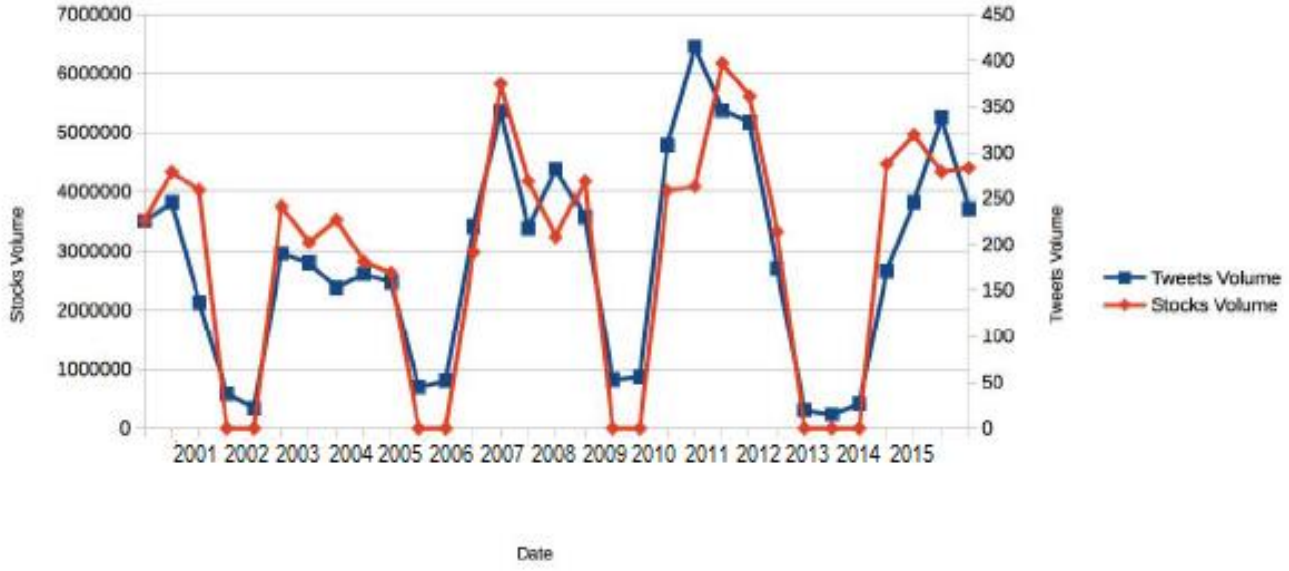
Only the first quarter prediction coincides rest are nowhere near the original stock value. The period between 2002 to 2006 is the only period predicted right. Efficiency for J.P. Morgan Chase given by the model is 23% with regularization coefficient of 0.5. Other values of regularization coefficient could have produced better results.

Microsoft:



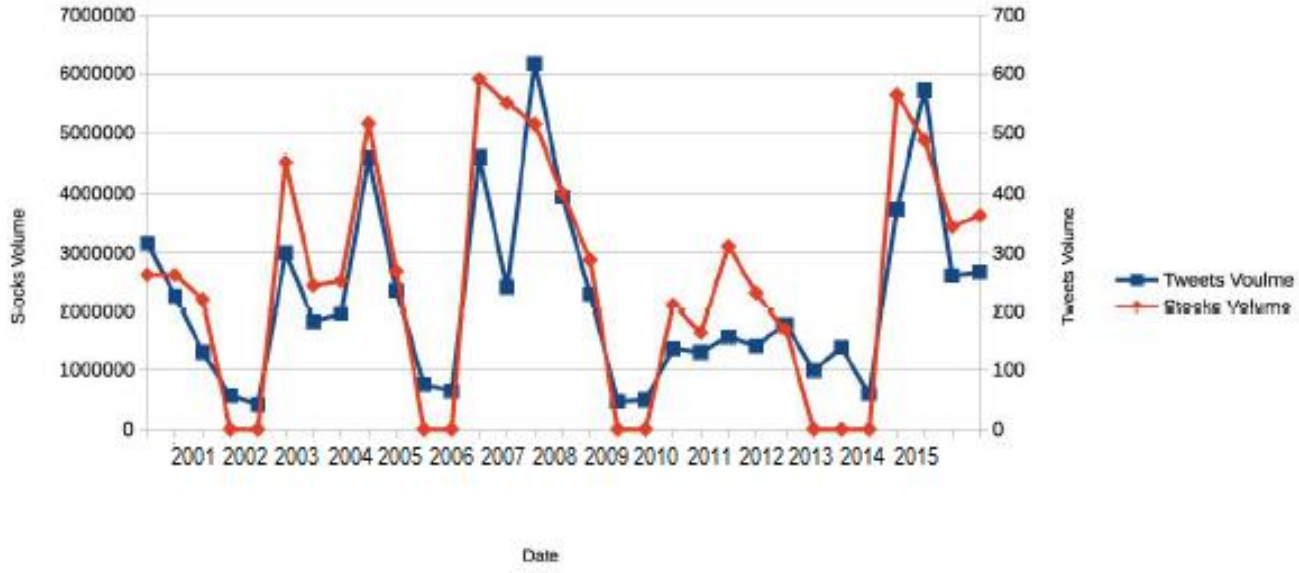
More or less a good fit with an efficiency of 72%. With certain deviations during the years 2009, 2010, 2011 and 2013. For the rest of the time span the lines seem to be lie approximately close to each other. This model had a regularization coefficient of 0.5. Other values of regularization coefficient could have produced better results.

Netflix:



All the quarter predictions seem to be in line with slight deviations here and there like for the years between 2005 and 2006 and 2010 and 2011. Netflix has got a high efficiency from this model with regularization coefficient of 0.5; about 91%. Other values of regularization coefficient could have produced better results.

Samsung:



With an average efficiency of 74% this model could predict stock value more or less accurately for Samsung as well. Slight deviations being in the year 2007 and from 2011 to 2014. This model had a regularization coefficient of 0.5. Other values of regularization coefficient could have produced better results.

Chapter 6

Conclusions

In this thesis, we analysed Twitter dataset to investigate if there is any correlation between tweets and fluctuation of the stock market. We analysed a much bigger dataset compared to the studies done in the past, selecting top ten most tweeted stocks for analysis, ranging from various domains. Out of total days for which we performed our experiment, we find out that for 78.3 % of days the values are positively correlated and the average distance between stocks volume and tweets volume, on correlated days, is lower than 16%. We also performed sentiment analysis on the tweets to find correlation among the sentiment and fluctuation in the stock price. To understand the correlation between the daily sentiment about one specific stock and the performance of that stock in the markets we performed sentiment analysis on the tweets. As a supplement, we also analyzed the network structure of the dataset to understand the relation among users of the Twitter dataset.

6.1 Future Work

We are planning to extend this work in following dimensions. One by using other regularization constants over the dataset. Next important line of work is to analyse richer dataset. By richer, we mean
following

- By clubbing web data (for example, from blogs) with Twitter dataset.
- Dataset which spans across over a much longer period of time.
- By considering not only cashtagged tweets, however also the one which uses the name of the companies in their tweets.

In this thesis, we only performed experiments on ten top most tweeted stocks. However, we would like to perform similar experiments on a larger number of stocks. Also, we would like to perform experiments on larger number of months.

We would also like to propose a model based on our rich dataset which can predict the stock market in the future.

Our sentiment analysis method being used is just a first attempt; we can call it Term Weight Analysis or Vocabulary Analysis. In future we would like to improve our sentiment analysis based on techniques such as the Natural Language Processing or other techniques rooted in machine learning, using Support Vector Machines.

Chapter 7

References

[1] https://about.twitter.com/company

[2] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment,” 2010.

[3] S. Kampakis and A. Adamides, “Using Twitter to predict football outcomes,” Nov. 2014.

[4] L. Zhang, “Sentiment analysis on twitter with stock price and significant keyword correlation,” 2013.

[5] J. Smailovic, M. Grcar, N. Lavrac, and M. Žnidaršic, “Predictive sentiment analysis of tweets: A stock market application,” in Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, vol. 7947, pp. 77–88, Springer Berlin Heidelberg, 2013.

[6] T. R. H. of Twitter, “http://www.businessinsider.com/how-twitter-wasfounded-2011-4,”