Stock Market Analysis using Textual Data

by

Jayesh Dhoot

A Project Report Submitted
in
Partial Fulfillment of the
Requirements for the Degree of
Master of Science
in
Computer Science

Supervised by

Dr. Carol Romanowski

Department of Computer Science

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

May 2016
Approval

Name: Jayesh Dhoot

Degree: Masters in Computer Science

Project: Stock Market Analysis using Textual Data

Advisor: Dr. Carol Romanowski
Associate Professor

Signed: 

Date: 
I would like to thank my advisor, Dr. Carol Romanowski, for the guidance and advice she has provided during the duration of project. I have been lucky to have an advisor who has always responded to my questions immediately. I would also like to thank Computer Science Department staff at Rochester Institute of Technology for their support. Finally I would like to express my gratitude to all my friends who helped me with their feedbacks.
Abstract

Stock Market Analysis using Textual Data

Jayesh Dhoot

Supervising Professor: Dr. Carol Romanowski

This paper tells about using sentiment analysis over textual data about companies and using a supervised learning model to see the impact that they can have. Due to the profit that may be involved in the stock market in selling and buying stocks big financial companies have been investing sentiment analysis of the textual data that may be available about different companies. We take into consideration different news articles written by established writers and also data from Twitter relating to the stock market specific companies. We then evaluate all the data with respect to a company to be able to advice if the company might have a positive impact in the near future.
# Contents

Approval .......................................................... ii

Acknowledgments ................................................. iii

Abstract ........................................................ iv

1 Introduction ................................................... 1
   1.1 Problem Statement .................................... 1
   1.2 Motivation ............................................ 2

2 Related Work .................................................. 4

3 Implementation ............................................... 6
   3.1 Data Collection ....................................... 6
      3.1.1 Sample Data ..................................... 7
   3.2 Data Cleaning ......................................... 8
   3.3 Training Data ......................................... 8
   3.4 Supervised Learning .................................. 8
      3.4.1 Classification .................................... 9
      3.4.2 Regression ....................................... 9
   3.5 Algorithms Used ..................................... 9
      3.5.1 Bernoulli Naive Bayes Classifier ............... 9
      3.5.2 Multinomial Naive Bayes Classifier .......... 10
      3.5.3 Logistic Regression ............................. 10
      3.5.4 Support Vector Machine ....................... 11
   3.6 The voting method for sentiment analysis .......... 11
      3.6.1 Case 1: ........................................... 12
      3.6.2 Case 2: ........................................... 12
      3.6.3 Case 3: ........................................... 12
   3.7 The Flow Diagram .................................... 13
# Results

4.1 Error rate ................................. 14
4.2 Sample Results ............................ 15
   4.2.1 Apple ............................... 15
   4.2.2 Alphabet .............................. 16

# Conclusion and Future Work .......................... 18
5.1 Conclusion ................................ 18
5.2 Future Work ................................. 19

Bibliography ........................................ 20
List of Tables

3.1 Input Table ............................................................. 13
4.1 Accuracy Table ......................................................... 14
List of Figures

3.1 Sample Tweet Data ............................................. 7
3.2 Sample Article Data ............................................. 7
3.3 Flow Diagram ..................................................... 13
4.1 Apple stock prices ............................................. 15
4.2 Apple sentiment analysis ....................................... 16
4.3 Alphabet stock prices .......................................... 16
4.4 Alphabet sentiment analysis ................................... 17
Chapter 1

Introduction

1.1 Problem Statement

Stock prices and their forecast is very important in the global market as well as to individual investors. However, as we know it indeed is a very challenging job to have a model which may rightly predict the future of the stocks and their price movements. The main reason for this being such a difficult job is the number of factors that may be involved in calculating the movements. The company history, recent movements in the competitors business, actual price of the stocks, the CEO of the company, the feel the people might have about a company, and the global economy and how analysts predict what should be expected are just some of the things that we might have to be considered for a perfect prediction model.

There is a popular phrase called the wisdom of the crowd which I would like to mention here. It basically is the judgement that the crowd provides over each individual judgement. What happens is the noise that might have come from each individual person in cancelled as a reason of the crowd answer. Many a times the whole crowd maybe led into thinking positively or negatively about a certain company and that could well result into a profit / loss for a company. The important thing that I do want to mention here is the impact that such a crowd could have on a companys future is astounding. This in fact does occur enough times for us to know what it means.

Out of all those things two things that do stand out are of course the stock prices and the news that is available about the companies on various social and business platforms. There have been different algorithms which take into consideration the stock prices and the
related movements and the history of the stock price fluctuations in similar situations. The downside of such algorithms is that they do not consider the public sentiment or what the public thinks about the company. The impact of the sentiment that is amongst the public is underrated in these algorithms.

There are two normally used categories of stock market analysis, technical and fundamental [1]. Technical analysis analyzes future prices based on the historic price movements whereas fundamental analysis deals with analyzing the stock market based on the other hand looks at economic factors which can be called as the fundamentals or the essentials.

Like some other future prediction algorithms, Technical Analysis will not generally give an exact detail of things that may be happening but will be able to tell what is likely to happen in the time to come based on what has happened previously in similar situations. The prices finally do fluctuate on the basis of the supply and demand and a few other market related values. They can normally have a look at data that lasts some weeks at the minimum which is quite different to that of fundamental analysis. Technical Analysts believe that the economic factors that fundamental analysts care about are already taken care of in the stock prices and the various charts that they might create.

Fundamental analysis analyses the market based on the economic factors [2]. They look into the depths of the company. Things like the revenue, the liabilities that they may face, the share they own the market etc. Most of the part in such analysis in the early days is spent on learning about the balance sheets, income statements etc.

In this project, I applied text mining algorithms to news articles and twitter data about companies which give us information where and how the company has been in news recently. This data is then used to analyze the general sentiment about the company and give us an overall sentiment based on all the recent articles which should be considered.

1.2 Motivation

There were two important motivations for doing this project:
• **Importance of Stock Market.**

The stock market can have a big impact on the economy of a country. If the share market has a steep fall due to any reason then it could possibly be a sign for a recession. The most prominent example being the great recession of the 1930s which was caused by the stock market crash a year earlier [7]. It is not always the case that a stock market crash is followed by a recession but it is every once in a while.

• **To show how useful could unstructured data be.**

The world today has a majority of its data in an unstructured format. Due to the large boom in the number of smart devices that are being used by an individual, the amount of data generated by them is also huge. Some studies say that the amount of unstructured data maybe in the range of 70-80% overall [3]. Through the project I would again want to stress the importance that such kind of data might have with the stock market being my way to take on this issue which I feel would be even more important going forward.
Chapter 2

Related Work

Robert P. Schumaker and Hsinchun Chen made a new machine learning model used for prediction for news articles which used different natural language processing techniques like using the bag of words, Noun Phrases, Named Entities and had a look at a huge number of articles and quotes which were related to the stock market [9].

Shah and Mohri, made another machine learning algorithm this time based on linear regression and support vector machines using online learning. They also provided the advantages and disadvantages of each method that they used. The machine learning method hinted generally at the text mining [4].

Nan Li and Desheng Dash Wu, text sentiment analysis in which they studied online forums which were present on financial sites. They used emotional analysis on the data for their study. It was totally different from what others were doing in that period [6].

Xiangyu Tang and Chunyu Yang, Jie Zhou made an algorithm which used news articles along with some time series analysis to give forecasting results. This was the first time when the idea of text mining with time series analysis was used by anyone [11].

Antonina Kloptchenko, Tomas Eklund, Barbro Back, Jonas Karlsson, Hannu Vanharanta and Ari Visa made research based on the data mining so that they could have an insight into the company future by having a look at the company's quantitative and qualitative part of
their reports. The drawback of their report was that it could be applied only on the annual report of the company and couldn't have worked on any other data that the company might have [5].

Marc-Andr Mittermayer and Gerhard F. Knolmayer applied text mining approaches and had a survey on different approaches that could have been possible. They compared eight different systems with one another and provided results for the same. The only problem with them was none showed any possible decision making possibilities [8].
Chapter 3

Implementation

The start of the research was some of the related work that has been mentioned earlier in the paper. Here ahead, the research that I did and the method followed would be mentioned. The process began with collecting the most recent news articles and tweets related to individual companies and storing them together so that they could be used while calculating the overall sentiment for that company. Cleaning the data is the next step and one of the most tedious steps out of all. This was followed by training the algorithms with a training data set and then using the new model on the data that was being corrected and finally predicting the sentiment of all the data individually. The final part of the project would take all the advice related to the company and apply a simple algorithm to see in which direction are things going on with the company.

3.1 Data Collection

The major source of data that was being collected was Bloomberg.com, The Wall Street Journal, Yahoo! Finance and Twitter. The main reason these sites were to be considered was to be sure that the source of the data was genuine.

Some of the articles from the websites had to be collected manually as there was no way to be sure if the article that was being collected was of any importance or not to any of the companies that were present.

The data from twitter was collected using a python API called tweepy and also from a website called the stocktwits.com. It provided us with all the tweets about a company
within the past 7 days with the stock symbol that was provided. All this data was saved in a text file to be processed further.

3.1.1 Sample Data

Here is a sample tweet and article data

![Sample Tweet Data](https://t.co/rK1NVDk0T0)

Figure 3.1: Sample Tweet Data

Figure 3.1 shows a sample tweet that was collected from twitter about $AAPL that is the stock market symbol for Apple Inc. The instant reaction after reading this tweet of course is negative. We collected a bunch of tweets for each company and had that separated by line which would lead to easy processing in the further stages.

Figure 3.2 shows an article talking about Under Armour, a sportswear brand. It mentions a variety of things almost all of which are positive. It talks about the recent success of their leading ambassadors and how that has affected in their brand gaining price in the stock market which obviously is a positive thing.

![Sample Article Data](https://t.co/rK1NVDk0T0)

Figure 3.2: Sample Article Data
3.2 Data Cleaning

This was one of the most tedious jobs at hand. Collecting data was interesting in a way as it had me look through various websites for data that could have been useful and then getting to know more about the kind of data that have been doing well recently.

The cleaning for articles was basically converting the HTML files to text documents which could make the further processing very easy. For converting from HTML to text we had to get rid of all the HTML tags and then just get the plain text in a file. Initially I had a small code written in python whose NLTK library provides a very easy way to convert the HTML document into a text document.

The cleaned data as expected would have no tags and would also get rid of any links that may have been there in the text. The link part was especially important for the data from twitter as many users tend to have links to various articles present there.

3.3 Training Data

The training data was partially created manually and partially got from an on-line source. The training data for the data acquired for twitter was provided on-line whereas the one for the articles was created manually.

For the manual process, I had to read the articles myself and then assign a sentiment to each of the articles. This was needed as there was a big difference in an individual instance of the training data that I had and the articles size. The training data was a good source for the tweets that I had.

3.4 Supervised Learning

Supervised Learning is a machine learning technique that allows to make the machine learn through previous examples in similar situations. So suppose you want to predict if a certain team will win the world cup that is taking place this year. So you would have a look at the data for the previous countries which actually won the world cup and some reasons that
may have contributed for their success. The reasons could be their long run of wins or their rankings that resemble their dominance or that they are unbeatable on their home ground which is where the cup is hosted etc. It could also well be a combination of all these things with some others as well.

So the model that is created will now see if any of the countries that are participating in the world cup do resemble any of the similar processes as of the previous winners. The ones that come closest to the earlier models will be the winners according to the model.

So the model basically learns through the previous results and then applies the knowledge it has gained into the new data that is coming through to predict what it thinks.

Supervised learning could be divided into two categories:

3.4.1 Classification

In this type of learning the goal is to classify the given data into one of the already present classes. It could be positive negative, span not spam etc. Predicting who will win the world cup is a classification problem.

3.4.2 Regression

This type of learning is useful for predicting continuous observations. It normally is used to predict values which are real numbers. One of the applications here is calculating the stock prices that may occur in the future.

3.5 Algorithms Used

3.5.1 Bernoulli Naive Bayes Classifier

It uses the Bayes probability theorem which is reputed for creating clean and simple classifiers. This classifier uses the theorem with independence among the features. The Bernoulli
Naive Bayes classifier is extensively used in the categories of text retrieval and categorization. It has boolean input values. This classifier makes use of term occurrence features instead of term frequencies as opposed to the tf-idf concept. The equation for this classifier gives the probability of a class that generates a particular term. The Bernoulli Naive Bayes classifier is not the same as multinomial and works well with absence of terms classification. It is particularly useful with short text documents.

### 3.5.2 Multinomial Naive Bayes Classifier

Multinomial Naive Bayes Classifier is a variation of Naive Bayes that works well with larger text documents. This algorithm provides more information than just the term occurrence. It represents the frequencies of the occurrence of certain events with multinomial probabilities. This algorithm is a generalized form of the Bernoulli Naive Bayes Classifier. It takes into account the conditional probability of a given term. This classifier is faced with several problems and is not competitive enough with SVM.

### 3.5.3 Logistic Regression

Logistic regression is a regression model where the dependent variable is categorical. It is a direct probability model. The data-sets in logistic regression are one or more independent variables that determine the outcome as opposed to linear regression. The output measured is a dichotomous variable. This regression is used to predict discrete values. A sigmoid function is used to fit the output in $[0,1]$. If the output is less than 0.5 then it belongs to one class and if it is greater than 0.5 it belongs to another class. The parameters can be learned using the Maximum Likelihood Estimates approach. The log likelihood of $w$ is computed and gradient descent is used for learning the parameters. While computing the update the equation consists of the parameter eta which can be tricky to set. Another way to speed up convergence is using Newton’s method where we compute the Hessian. The Hessian is a second order derivative. A Gaussian prior is added while performing the regularization with logistic regression since over-fitting is an issue with large numbers.
Bayesian Logistic Regression uses Laplace approximation to compute the posterior for w.
Logistic regression resolves the categorical outcome variable that violates the assumption
in normal regression about linearity. It resolves this concern by algorithmic transformation
on the outcome variable that permits us to model a nonlinear association in a linear way.

3.5.4 Support Vector Machine

Support vector machine is a classifier that uses a separating hyperplane for the discriminative analysis. SVM will give an output in the form of a hyperplane to classify the data-set. SVM is similar to perceptrons. The SVM learning task is an optimization problem deals with finding weight vector w and bias b that ideally give zero training error. An SVM uses the radial basis function (RBF) kernel which has a bell curve. SVC is another implementation of the same algorithm. linearSVC is an implementation of scikit-learn which performs multiclass classification on a given dataset. LinearSVC is based on liblinear in scikit-learn and only supports a linear kernel.

This class is able to support both dense as well as sparse input and the multi-class support is handled according to a one-vs-the-rest scheme [10].

Since they are both classifiers they can extensively be used in the field of text mining.

3.6 The voting method for sentiment analysis

So as it can be seen in the model we trained 5 different algorithms described earlier for finding the sentiment of the data that is sent to our model. The main reason for using five different algorithms is that we can have more accuracy in our model as we take inputs from all the algorithms and then give a result combined from all their outputs.

So each individual algorithm gives us the confidence it has along with the classification it results into based on the training data that was provided by us. We then combine the results of all the algorithms and take into consideration the overall classifications. The confidence is not considered while the initial classification in being done. If 4 out of the
5 algorithms are giving us a positive or a negative answer we take that to be a confident answer and then use it.

Let's say for example the algorithms return the following answer for a given text input Table 3.1. There are a few cases in which it needed to be decided how I wanted to combine the results of all the algorithms together.

As the algorithm is generated in this way, we will just have 3 possible cases. The cases where two algorithms say positive or one algorithm says positive are equivalent to saying three algorithms are saying or four algorithms are saying negative respectively which have already been covered.

### 3.6.1 Case 1:

All the 5 algorithms say that the input text is positive or negative. In that case we have no doubt about the sentiment of the text and we confirm that the sentiment is what all the algorithms are saying. The confidence is such circumstances is 100%.

### 3.6.2 Case 2:

Four algorithms say that the text is positive and one says that it is negative. In this case we still go with the majority of the results of the algorithms. So the answer for the input data that presents us the above case would be positive along with a confidence of 80% that is equivalent to (4/5).

### 3.6.3 Case 3:

Three algorithms say that the input text is positive and 2 say that the input text is negative. In such situations we need to consider the fact that none of our algorithms give us the perfect accuracy. There is always a chance of a false positive here or there. So in this situation we still give the answer as positive for the text but here is where the confidence comes into picture. We keep the confidence to 60%. The confidence comes into the picture only when
<table>
<thead>
<tr>
<th>Algorithm used</th>
<th>Classification</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nave Bayes</td>
<td>Positive</td>
<td>75%</td>
</tr>
<tr>
<td>Multinomial Nave Bayes</td>
<td>Positive</td>
<td>87%</td>
</tr>
<tr>
<td>Bernoulli Nave Bayes</td>
<td>Positive</td>
<td>69%</td>
</tr>
<tr>
<td>Linear support vector classifier</td>
<td>Negative</td>
<td>76%</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Positive</td>
<td>77%</td>
</tr>
</tbody>
</table>

Table 3.1: Input Table

it is 60%. While plotting the graph we do not consider any data which has sentiment value less than 80 which is equivalent to saying if it gets less than 4 votes.

### 3.7 The Flow Diagram

So The Flow Diagram:

![Figure 3.3: Flow Diagram](image-url)
Chapter 4

Results

The results achieved for the overall sentiment module of the project would be in the lower seventies. This result as achieved using the combination of the results for all the algorithms together. Table 4.1 shows the accuracy available for all the voting algorithms.

So combining the results of all these would give an approximate accuracy of 70.8%. Yes the overall accuracy as expected does go down because we combine the algorithms with each other. But the overall results for the project do have a good result as we use the output from 4 out of the 5 algorithms to conclude if the given code is positive or not.

4.1 Error rate

The overall accuracy of the code had one important standpoint.

- The use of four out of five algorithms made it less error prone for false positives.

Many a times when the algorithm gave values which were 3-2 in favor of one of

<table>
<thead>
<tr>
<th>Algorithm used</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nave Bayes</td>
<td>72%</td>
</tr>
<tr>
<td>Multinomial Nave Bayes</td>
<td>70%</td>
</tr>
<tr>
<td>Bernoulli Nave Bayes</td>
<td>69%</td>
</tr>
<tr>
<td>Linear support vector classifier</td>
<td>73%</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>71%</td>
</tr>
</tbody>
</table>

Table 4.1: Accuracy Table
the sentiments it was the case that the algorithms which was supporting the right sentiment had the smaller number. So in this case we just did not consider the result instead of considering the wrong result. As the amount of input data was big enough, we could afford to do something like this without having any issues with respect to the availability of data.

4.2 Sample Results

4.2.1 Apple

Figure 4.1 is the graph for the Apple stock prices for the week of 2\textsuperscript{nd} May with the sentiment values for Apple, in Figure 4.2, for the same amount of time. As we can see the sentiment value for the stock goes almost in the same direction as the actual stock price for the stock. Even though the graph is pointy in most of the places instead of the curves that may be.

One of the things to be noticed in the graph is the sentiment data value on the x axis. The value crossed 100 (articles + tweets) for a period of 5 days. The total number being more than 15-20 articles per day gave us an actual insight of the data rather than having 2-4 articles per day which would not give us the actual values we were looking for.

![Apple Stock](image)

Figure 4.1: Apple stock prices
Figure 4.2: Apple sentiment analysis

### 4.2.2 Alphabet

Figure 4.3 is the graph for the Alphabet stock prices with the sentiment values for Alphabet in 4.4. As we can see the sentiment value for the stock goes almost in the same direction as the actual stock price for the stock. Even though the graph is pointy in most of the places instead of the curves that may be.

Figure 4.3: Alphabet stock prices
Figure 4.4: Alphabet sentiment analysis
Chapter 5

Conclusion and Future Work

5.1 Conclusion

- There is a strong connection between the news articles or company news and the movement of the stock prices. This can be determined by the similarity in the two graphs that have been shown. The stock market values fluctuate according to the fluctuation in the sentiment of the people.

- A decent amount of data is needed to produce a result which matches closely to the actual results. Some of the companies that I had, did have a shortage of data that could be used for sentiment analysis for the project. For example, Intel had less than 10 tweets in a period I was checking for tweets for some of the big companies. A drawback is that not enough data is available for small companies or companies that have been around for a few years but still did not have any mentions in this post.

- It could also be used with the actual stock prices to get more detailed information about the company. Stock prices and sentiment analysis together could be a very great force to reckon with in the market of Stock Market Analysis. It would be a combination of Technical and Fundamental Analysis together. This hasn't been proven yet but it has been successfully used by some to get some great results.
5.2 Future Work

- The accuracy could be increased with a better training set. If the training set could be an actual stock market related set then I'm sure the accuracy of this code would have been much better than what it is right now.

- The accuracy could be increased with a better training set. If the training set could be an actual stock market related set then I'm sure the accuracy of this code would have been much better than what it is right now.

- Combining with the actual prices of stock will help in gaining an insight about possible stock prices as well.
Bibliography


