Analysing News for Stock Market Prediction

V. V. Ramalingam¹, A. Pandian², shivam Dwivedi³ and Jigar P. Bhatt⁴
Department of Computer Science and Engineering, S.R.M. Institute of Science and Technology, Chennai, India
ramabi1976@gmail.com

Abstract. Stock market means the aggregation of all sellers and buyers of stocks representing their ownership claims on the business. To be completely absolute about the investment on these stocks, proper knowledge about them as well as their pricing, for both present and future is very essential. Large amount of data is collected and parsed to obtain this essential information regarding the fluctuations in the stock market. This data can be any news or public opinions in general. Recently, many methods have been used, especially big unstructured data methods to predict the stock market values. We introduce another method of focusing on deriving the best statistical learning model for predicting the future values. The data set used is very large unstructured data collected from an online social platform, commonly known as Quindl. The data from this platform is then linked to a csv file and cleaned to obtain the essential information for stock market prediction. The method consists of carrying out the NLP (Natural Language Processing) of the data and then making it easier for the system to understand, finds and identifies the correlation in between this data and the stock market fluctuations. The model is implemented using Python Programming Language throughout the entire project to obtain flexibility and convenience of the system.

1. Introduction
Unstructured data has emerged as an efficient source for information extraction. Text mining through large databases to obtain the desired characteristic information is rapidly growing and the derived information can be used for stocks prediction. Although, these databases are very large and the data collected is not only from the news but also general opinions on the internet. We propose a method to collect online data from a social media platform, commonly known as Quindl. The size of data is very large and therefore the data is linked to form a csv (Comma Separated Value) file and is then parsed to obtain cleaned data, which is the required characteristic data for predicting the stock market values for the user. Sales, Marketing, Manufacturing are other areas where this data can be used. Since the data collected from Quindl also includes public opinions, it cannot be completely relied on. The data is presented or interpreted differently by different person. Therefore cleaning this data is very essential. The news presents both the positive and negative aspects in a neutral tone. Another important aspect is the correlation coefficient which is nothing but a statistical measure reflecting correlation between two securities. This coefficient indicator helps investors diversify by identifying securities with a low or negative correlation to the stock market. The indicator is negative when the securities move in opposite direction and vice versa.
2. Problem Statement
In particular, the prediction of stock values using news or personal opinions on the web constitutes an attempt to extract meaningful characteristics from texts using text mining. However, it is still difficult to analyse and extract meaning because such text data comes in many diverse types, and comprises multi-meanings, complex words and various features. The news, generally, presents both positive and negative aspects of the stock markets in a somewhat neutral tone, making it difficult to identify the underlying truth behind such news. Furthermore, there is a danger that the news may be analysed and interpreted differently depending on the writer.

3. Related Work
3.1. Opinion Mining
It is referred to as computational techniques used for extracting useful characteristic data from public opinions on social media platforms or from the news. Automatic movie review mining or summarizations of movies are some of the works showing the effectiveness of opinion mining. Using the same technique the market moods monitored from Quindl can be used to predict the flow of stock prices. This makes the extraction of required data from public opinions easier and helps the user in determining the stock market values to ensure safe investments. To obtain better results it is important to build a lexicon of opinion mining.

3.2. Stock Prediction using the news
The relationship between news and stock prices has been in research for quite some time. The web has emerged as a great source of financial information analysing which can aid the stocks prediction. Various stock prediction systems collect and analyse this news to predict the fluctuations in stock market. This has emerged with the rapid growth of unstructured big-data over the internet. The financial message boards are scanned and analysed extracting the sentiments of the authors followed by identifying their correlation with the stock values. The data collected is large unstructured data from the news where the required characteristic data is hidden and therefore this data first needs to be cleaned and then identifying its correlation coefficient helps in predicting the fluctuations of the stock market and also checks the model accuracy.

![Figure 1. System Architecture](image-url)
News gathering is the first step of obtaining our data from Quindl. This data is stored in a database. The next step is Natural Language Processing in order to extract sentiments from the unstructured data. The sentiment dictionary is created which is in form of a csv file. This data is then used for sentiment analysis and opinion mining followed by supervised learning experiments aimed at predicting the stock market fluctuations.

3.2.1 Data Collection
This process involves collection of big, unstructured data from news or any social media platform, Quindl in our case. The data over some specific period of time is collected through text mining to obtain the required data only. This data is still bid unstructured data and needs to be cleaned to use for our stock market predictions. The data collected over such platforms has to be approved from the users entering this data over Quindl. Therefore, only the data permitted to be accessed is used in the model. Being in the category of public opinions this data is quite big and in unstructured form.

3.2.2 Sentiment Analysis
The sentiment dictionary plays a crucial role in opinion mining, essentially to build in linguistic resources for the data. Essentially this is done to classify the data, obtain the comma separated values file. This classification helps in narrowing down the work for cleaning the data, acquiring the essential characteristic data to be used for our model. To develop this dictionary, is basically in two phases. The first phase is to delete the punctuations, numbers, English one character words etc. whereas the second phase involves forming an opinion of the data calculating the recurrence probability.

3.2.3 Stock Predictions
The sentiment analysis helps in deciding a sentiment score also known as the correlation coefficient indicator to test the accuracy of the model and also determine the stock market fluctuations. The cleaned data from the unstructured big-data is converted to opinions suitable for stock market predictions. As is already known, the stock market can either fall or rise any day. The training dataset as well as the validation dataset helps in determining the accuracy of the model.

4. Results and Discussions

4.1 Load Modules
Firstly we load the appropriate modules into our python environment. This can be done using import function for including appropriate and required modules from the library onto our environment. There are many modules we will be using in our model, one of them being matplotlib which is basically for plotting the statistical data. This representation makes it easier for determining the market fluctuations and predicts the values.

4.2 Get Data

4.2.1 Collecting Data
Collecting the data set required for our analysis and predictions should be reliable and accurate. For this purpose only we collected S&P Stock Index data at any particular frequency of any time period, as desired by the user for any finance company. This data is reliable and used in our analysis for determining the fluctuations in the stock market.

4.2.2 Cleaning Data
This implies obtaining the relevant and required information from our big unstructured data file which can be used by our model. This is a very important part for our model. The obtained data is now stored in the form of a .csv file which separates the value through comma and represents it in column form. This makes it easier and faster for the machine to analyze the data set.
4.2.3 Loading Data

Now we need to load this data into our python environment. Here we can specify the range of our data set i.e. the time period over which we want to collect out data set.

4.3 Plotting our Time Series

Plotting the data makes it easier to represent and analyze the data over a particular range, as desired by the user. Also, the comparison of the predicted values with the actual values is easier from the representation. Regression Analysis is required for this part as it helps it obtaining multiple data sets from the internet, analyses them and plots them for representation.

**Stepwise Regression Analysis**- It is the name given to iterative multiple regression models. We are using regression technique as we want to add and compare multiple data sets for our graph representation. Also many other regression techniques are available and can be used as per the choice of the user. For example, Consider an investor who decides to imply more weightage (importance) to the recent years of dataset from the whole range of dataset represented. In this case, weighted least squares regression analysis is used. The quandl integration, in our case, turns out to be a reliable source of data as it indexes millions of economics, finances datasets. Also, the use of Matplotlib makes it easier to pot multiple regressions on one single graph.

A representation of S&P 500 stock Time series as obtained is shown below. The curves represent the different values used in regression over a specific time range.

![Time series plot for S&P 500](image)

**Figure 2.** Time series plot for S&P 500

4.4. Model Estimation

The Correlation function now comes into play. This function is used for determining the accuracy of our model. Since we have already plotted the time series graph determining the correlation function becomes easier and faster. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are two different function used, both are predicted and represented in different graphs.

The ACF and PACF plots for our specified data set are represented below. Transforming the data in these plots is also very necessary to prevent the factor of time decay for our specified dataset. Since, each function uses its individual function over the dataset the representation of both graphs is different over the same dataset during the same time period. These functions help us in depicting the decay rate of our time series plot as well as in determining the values for the stock. Both these functions are thus necessary individually. Also the correlation function is the one which is used in determining the accuracy of this model.
Transforming our Data

This is generally done for non-stationary data. This is the kind of data having their distributions change over time i.e. the data with flexible values over the time range. Our dataset falls in this category, as we are taking up different values of stocks over a specified time range. The graph obtained after this transformation sets up an average value and represents both the high and low points (positive and negative) depending upon their average values. This turns out to be very useful when determining the fluctuations of any particular stock over specified time range.

A representation depicting a first difference time series plot i.e. after our data transformation is shown below.

Figure 5. First Difference Time Series Plot
This graph represents the same dataset which we used for obtaining the previous graphs already shown above. Such representations make it easier to focus on any one particular data over the specified time range through which analyzing the fluctuations as well as predicting the value becomes much more convenient and a faster process.

4.5 Evaluation

The statistical measures are used to test the accuracy of the system in predicting the fluctuations of the stock market. Recall and Precision are two such statistical measures which help in determining the quality of the results. Both these compare two standard performance results. In general, there is an inverse relation in between both precision and recall. An ideal model would have both high recall as well as precision.

4.6 Analyzing Sentiment

The sentiment dictionary being an essential factor for opinion mining has to be created for making the mining of the unstructured data easier. Developing a stock domain specific dictionary requires performing of NLP over the data collected. A sentiment score is then assigned to the words forming the dictionary. These words are the ones commonly seen as the headings on our csv file. Mining using these words, transforms the malformed csv file into a more structured data file involving the essential characteristics data required for our predictions.

4.7 Predicting Increase/Decrease in Stock Prices

The main aim of the model is to determine the correlation between the news sentiment and the fluctuations of stock market and also compare the results with the actual trend happening to test accuracy of the model.

4.8 Result Forecasting

Now the final step after we have plotted our values is to compare it with the actual values of the stock market. A method known as predict () is used for this step. Once we obtain our final graph, to compare it with the actual values we need to merge them both. The regression function also comes into play for obtaining the specified values over a specified time period preventing any decay in any of our data represented continually to obtain even the attest values.
As depicted in the graph above, both the values for the actual as well as the forecasted values are merged together to form a single graph representation. To understand the comparison between the values, the user, over any time period within the time range can zoom in on the graph and only that particular part comparing the actual and forecasted values would be displayed. A representation of this is shown below.

5. Conclusion
Stock markets vary over different time periods and analyzing the news makes it easier for determining the trends of the market. Thus it can be said they have a relationship in between them. The model was developed based on this relationship and python language was chosen to make the model lesser complex. The accuracy of the model is also determined along with the forecasting results over the time period range. The data set is the main component and therefore should be from a reliable source keeping the latest values also in its record. Regression used for plotting this data also proved to be very effective. Python environment provides Matplotlib which makes it easier to plot different values over time ranges easily over a graph. Although many more systems are available for such predictions, lesser time and program complexity makes this model a step ahead of them. Also, quindl provides reliable and authentic data set.

References
[1] Pang B and Lee L 2008 Opinion Mining and Sentiment Analysis Foundations and Trends in Information Retrieval 2 pp1–2
[2] Schumaker R P and Chen H 2009 Textual Analysis of Stock Market Prediction using Breaking Financial News: The AZFinText System ACM Transactions on Information Systems 27
[3] Fu, Lee K, Sze D Chung F and Ng C 2008 Discovering the Correlation between Stock Time Series and Financial News Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology
[4] Chen H and Zimbra D 2010 AI and Opinion mining IEEE Intelligent Systems
[5] Bollen J, Mao H and Zeng X 2011 Quindl mood predicts the stock market Journal of Computational Science 2
[6] Kim Y, Kim N and Jeong S R 2012 Stock-index Invest Model using News Big Data Opinion Mining Journal of Intelligence and Information Systems 18
[7] Zhuang L, Jing F and Zhu X Y 2006 Movie Review Mining and Summarization Proceedings of the 15th ACM International Conference on Information and Knowledge Management
[8] Mittermayer M A and Knolmayer G F2006 News CATS: A News Categorization And Trading System Proceedings of the International Conference in Data Mining
[9] Ahn S and Cho S B 2010 Stock Prediction using News Text Mining and Time Series Analysis Proceedings of the KIISE 2010 conference, 37
[10] Sehgal V and Song C 2009 SOPS: Stock Prediction using Web Sentiment Seventh IEEE International Conference on Data Mining – Workshops
[11] Jones, Eric, Travis Oliphant and Pearu Peterson SciPy: Open source scientific tools for Python