STOCK MARKET CORRELATION WITH PUBLIC SENTIMENT

Yajash Pandey
NYU, MSIS
Yp1113@nyu.edu

Raj Sekhar Bandakavi
NYU, MSIS
Rsb503@nyu.edu

Abstract- The project focusses on finding the correlation between public sentiment and stock price fluctuations. We have used Twitter data to get the public sentiment. We calculated the correlation of many stocks and found out that certain types of stocks have better correlation with twitter sentiment.

Keywords—analytics, stock market, Pearson correlation, RDD, Stop Words, sentiment, linear regression

I. INTRODUCTION

The Stock market is know for being unpredictable and it has always generated great interest among general public. It is well known that the presence of a stock in major news affects its price in one way or the other. Also, it is theorized by Effective Market Hypothesis that the change in sentiment about a company in the general public affects the price of its stock. Twitter can be an indicator of both these factors. This project focuses on stock market prediction using historical prices of stocks in different sectors and their sentiment on twitter. We found the Pearson correlation between the twitter sentiment and stock price movement to understand if twitter if a reliable indicator of the fluctuation in a stock's price. To check the goodness of the analytic, we have calculated the sentiment of various stocks on days different from the ones used in the model and tried to predict the price using the correlation originally calculated from the model. The predicted price was compared with the actual price of the stock listed on the stock exchange.

II. MOTIVATION

Tremendous amounts of historical data related to the stock market is readily available. With the amount of data, there are a lot of hidden patterns which can be derived. The way the prices fluctuate and correlate with public sentiment is what we are trying to understand. These results when compared with stock markets results will help us estimate the accuracy of the analytic and also help us add hidden factors if any which can be used to get an even higher accuracy in future. Social media is gradually reflecting behaviour of other complex systems. Every single day large amounts of text are transmitted online using social media platforms. Though each tweet may not provide significant information, but the aggregation of these tweets and by applying the filter of words relevant to the analytic, an overall public sentiment can be obtained.

III. RELATED WORK

The role of data science comes into picture to analyse the hidden patterns in the date and reveal the insights. To solve the problems, we must look at a problem from different perspectives. It has to be seen with technical as well business perspective to be sure about different relevant approaches. One such approach is to evaluate the causal conclusions which happens due to hidden factors.: [3] Correlation between two quantities is another concept where we will be able to understand the dependency between two quantities, their interdependencies.: [3] A business problem will never come up in a neat and organized way. It always needs to be broken down into the underlying possible components and then needs to be analysed so that to guide us to the correct path. The context of the problem needs to be analysed before any analytics is done on the data. To predict the stock market, the ANN model has been popularly claimed to be a useful technique for stock index prediction because of its ability to capture subtle functional relationships among the empirical data even though the underlying relationships are unknown or hard to describe.: [2]. With many data sources one data source turns out be better than the other one. In our data sources we are looking into a specific set of input which can be used as a better indicator of the market.: [2]. The fundamental approach has been to get public opinion of a company. Sentiment analysis is done on the social media text which is generated every second in huge volume. The data is taken from social media using API's and crawlers in the data collection stage. Tweets consists of many acronyms, emoticons and unnecessary data like pictures and URL's. So, tweets are pre-processed to represent correct emotions of public. For pre-processing of tweets, a three-stage filtering has been employed: Tokenization, stop words removal and regex matching for removing special characters.: [5]. Also, to clean

the data, it is made sure that tweets are taken once that is retweets are not considered as they will give incorrect results. The words which are not adding to the analytic which means words which are neither positive nor negative can be removed. URL 's from the posts can be removed as they will not be used to analyse the sentiment. Cleaned data needs to be loaded into HDFS.: [4]. Also, to implement this a dictionary approach can be used where words are present with a score and tweets containing the words will be given a particular score. The data which is processed will give us three indicators which means that text can be classified as positive, neutral and negative. The sentiment of different social media is combined to give a total sentiment of the company. A biased opinion of a single individual is not required, in this case the average opinion is required that is why average sentiment needs to be considered.: [4].

IV. DESIGN

The design focusses on using Map Reduce for data profiling and data cleaning. After ETL has been done, analytic was done in hive to retrieve insights from the cleaned data set.

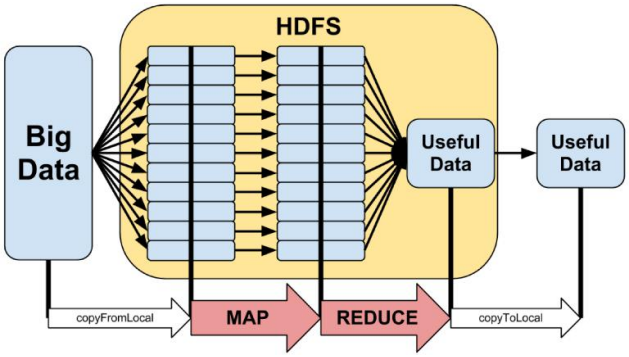

Fig. 1. Data profiling and Cleaning

Map Reduce is done in the above phases to get useful data from the big data in the data cleaning and profiling stage.

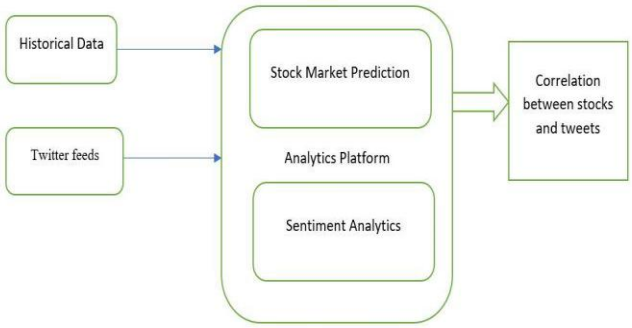

Fig. 2. Analytics model

Hive has been used on the useful data to get results for the analytic.

Twitter has made available 2 public API's to interact with its data i.e. tweets and several attributes about tweets. These are Search API and Streaming API. Streaming API has been used to filter tweets of a company continuously for our functionality. Using the official names of the company is a better approach to filter the tweets so that all irrelevant tweets which impact the accuracy of the analytic are removed in the data scraping process.

For twitter feeds, profiling code has been written to remove retweeted tweets and remove URL. These details need to be removed from the initial big data as they are not helping in the accuracy of the analytic and can be filtered out.

For cleaning the data, irrelevant tweets should be removed. Tweets are being loaded in csv format and extra commas which create difficulty while loading the table in hive need to be removed. Retweets should not be counted as those impact the overall public sentiment which leads to inaccuracy in the analytic.

For stock market analytic, the values required are obtained using Map reduce in the data profiling stage. In the data cleaning stage, we had checked for inconsistencies in the data as that can impact the accuracy of the analytic .The map reduce for cleaning was done on the profiled data to remove the inconsistencies in company name or ticket symbol for the company.

Stock Market Analytic helps us predict the change in closing price so as to accurately determine the stock price fluctuations. These part of the analytic helps us determine the stock price for a company in the future using the public sentiment and the correlation derived in the past for the company.

Sentiment Analysis will be done on the feeds using hive where tweets will be stored in a table and relevant fields will be selected from the tweets. Dictionary also needs to be stored in table where the tweets with words will be given a score and the average sentiment of the company will be calculated to get public opinion of the company.

There are two sets of output which are generated which are the stock market fluctuation and the twitter sentiment for a company. Both these outputs serve as input for getting the correlation for different companies. RDD is used to read the files and then the script is executed to determine the correlation coefficients between the same.

Pyspark script is written to perform correlation and used libraries such as Statistics and Multiples. The two parameters which are total sentiment and stock price fluctuation are passed as arrays and the correlation functions are called. Pearson correlation has been used which is a measure of linear correlation between two variables. The value of Pearson correlation lies between -1 and 1. A value of 1 implies that there is a perfect positive correlation between the two parameters. A value of −1 implies that the correlation between the two parameters is perfectly negative. A value of 0 implies that there is no linear correlation between the variables.

V. RESULTS

The analytic was able to get a positive correlation for companies in the financial sector. For example, the correlation of American Express Co was 0.339.The underlying reason for this is that the tweets about companies in the financial sector are more specific and are giving information which is reliable. For B2C (Business to Customer) companies, the correlation was not positive which implies that twitter sentiment is not a good indicator of their movement. The correlation for Ebay which is a B2C company was -0.443.This was due the fact that tweets about B2C companies were from a diverse sources and not completely relevant to the overall performance of the company.

VI.FUTURE WORK

More historical factors can be taken to derive the stock price in future using Machine Learning Models which will be taking price and behaviour of economic agents to improve the accuracy of the analytic.

Data dictionary which has been implemented in this project can be implemented using Machine learning model where it will learn and classify words as positive, negative and neutral.

Also, other correlation parameters which are not linear can be used to check for correlation between stock price indicators and public sentiments to give us more insights between other economic factors and sentiments affecting the stock market.

Getting tweets of companies for a longer period of time was a blocker as the twitter API helps us retrieve tweets for a period of 15 days. Third Party API where historical tweets can be retrieved help us get the tweets for a longer period of time, do our analysis on the tweets and understand the public sentiment of the company over a longer period of time.

The stock prices of a company can be easily obtained for a longer period of time and the derived correlation will be more accurate and can be tested as data for both sources will be available easily.

VII. CONCLUSION

The goal of the analytic was to check whether stock market movements are impacted by Twitter sentiment. We were able to understand that Twitter reflects the public sentiment of some companies. The accuracy of the analytic was higher for the companies in the financial sector where companies had a positive correlation with the public sentiment.

To verify the goodness of the analytic we performed the following steps: tweets of a company for a period of days are collected. The public sentiment of the company is calculated. Using this sentiment and previously calculated correlation coefficient, predict the stock price fluctuation. To confirm the analysis, check the stock price of the company on the stock exchange and verify if it is matching with the prediction.

V111. ACKNOWLEDGMENTS

We would like to express our gratitude to Professor Suzan McIntosh for introducing us to the world of Big Data and her guidance and support in the completion of the project. We would like to thank the HPC People for giving us the HPC Cluster to run our analytic. We would like to thank the Cloudera for providing the Cloudera VM to test the analytic.

REFERENCES

1. T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.

2. Predicting the direction of stock market index movement using an optimized artificial neural network model by Minogue qu, Yu Song

3. Data Science and its relationship to Big Data and Data Driven Decision Making by Foster Provost and Tom Fawcett

4. Stock Market Prediction: A Big Data Approach by Girija V Attigeri, Manohara Pai M M, Radhika M Pai and Aparna Nayak

5. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements by Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda and Babita M