Stock Market Prediction: A Big Data Approach

Girija V Attigeri , Manohara Pai M M, Radhika M Pai, Aparna Nayak
Manipal Institute of Technology,
Manipal- 576104, India

***Abstract*—The Stock market process is full of uncertainty and is affected by many factors. Hence the Stock market prediction is one of the important exertions in finance and business. There are two types of analysis possible for prediction, technical and fundamental. In this paper both technical and fundamental analysis are considered. Technical analysis is done using historical data of stock prices by applying machine learning and fundamental analysis is done using social media data by applying sentiment analysis. Social media data has high impact today than ever, it can aide in predicting the trend of the stock market. The method involves collecting news and social media data and extracting sentiments expressed by individual. Then the correlation between the sentiments and the stock values is analyzed. The learned model can then be used to make future predictions about stock values. It can be shown that this method is able to predict the sentiment and the stock performance and its recent news and social data are also closely correlated.**

***Index Terms* — Big data, prediction, social media analytics, machine learning**

I. INTRODUCTION

Stock market prediction is an important area of research and challenging aspect since many years. With an advent and aide of social media and technological advancements such as big data, it has created a wave amongst researchers.
Two basic principles of any financial derivatives are [1][2]

- Profit cannot be generated out of nothing
- No arbitrage principle: no opportunities for arbitrage that is there is no possibility of generating profit without any risk.

When developing a stock prediction model, concepts to be considered are

- Random walk theory: The theory that stock price changes have the same distribution and are independent of each other, so the past movement or trend of a stock price or market cannot be used to predict its future movement [1]. It is given by the formula

$$v(t) = v(t-1) + c(t)$$

$$\Delta v(t) = \frac{v(t) + d(t) - v(t-1)}{v(t)}$$

$$v(t): price\ of\ stock\ at\ time\ t$$
$$v(t-1): price\ of\ stock\ at\ time\ t-1$$
$$\Delta v(t): change\ in\ price\ of\ stock\ at\ time\ t$$
$$d(t): dividend\ at\ time\ t$$
$$c(t): adjustment\ term\ at\ time\ t$$

Since c(t) is the impact of all the publicly and privately available information on the stock, which makes prediction of $\Delta v(t)$ beforehand a difficult task.

- Efficient Market Hypothesis: It states that market price mirrors the assimilation of all the information available. As the new information enters the market the system immediately enters the unbalanced state and predicted correct change is eliminated by the new price. Hence given the information it is not possible to predict the future price of the stock [3]. However based on the information used to predict the future price EMH has three forms.
  - Weak form: Only the past information is considered.
  - Semi strong form: All publicly available information is used.
  - Strong form: All the information publicly and privately available information is used.

Stock market prices could be modelled using two approaches [4]. 1) Technical: Statistical analysis of the stock prices 2) Fundamental: Considers every detail available and behavior of economic agents that may affect price. It is performed on historical and present data, but with the objective of making financial forecasts.
In this paper we are using both technical and fundamental analysis. As depicted in Fig. 1. Technical analysis using machine learning algorithm on the stock market prices is done. Fundamental analysis using social media analytics is considered. Form of EMH for prediction is semi strong since only the publicly available information is used for prediction.

978-1-4799-8641-5/15/$31.00 © 2015 IEEE

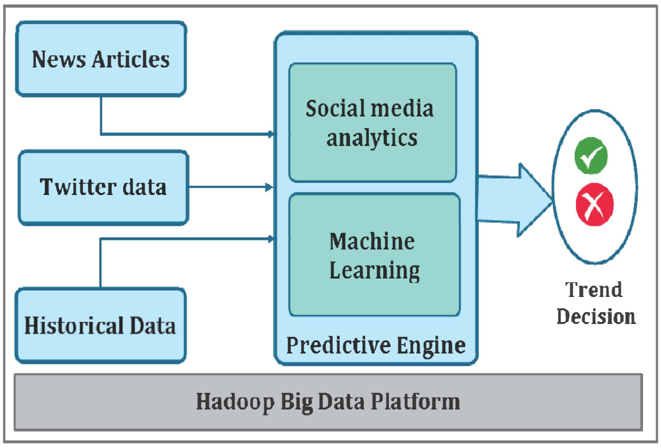

Fig. 1. Stock market prediction model

Machine learning is a well-established technique in a wide range of applications and has been broadly studied for its capacities in prediction of financial markets [5]. However inputs considered are mostly derived from the data within the stock market itself. Such separation might ignore the impacting factors on the price. Consideration to the publicly available data for stock market prediction can improve the accuracy. Social media analytics is powerful than ever. Social media refers to an informal, distributed content generation, propagation, and communication among various groups breaking the barriers of geography, society.

Data generated in social media is huge and unstructured and is generated every second. Technological support is very much essential to process such data in real time. Hence a big data approach is used for implementing a prediction model [8].

II. Literature Review

Many research groups are exploring stock market trend prediction using social media analytics. To detect he polarity of each tweet/news there are multiple methods.

i. Building own dictionary using semi supervised approach [11].

ii. Domain specific dictionary based approach.

In [11], semi supervised learning approach is used to build dictionary which is time consuming, because of initial level of manual work. Setting some threshold values words are added to either positive dictionary of negative dictionary. This approach is not suitable for real time analytics until the dictionary is complete. In [12], feedback of hotels from various web sites is analyzed using different open source tools on Hadoop. They have completely relied upon manually constructed dictionary, which is time consuming and we need to follow different dictionary for different products. In [13], only tweets are considered for analysis along with historical data. Stock market prediction in this model only considers open or close price of stock, even though high, low features are highly correlated.

Users are allowed to follow and comment on company's products or news in Twitter which is a micro blogging application[13]. As one of the most popular social media, more than millions users post over 500 million tweets every day. Much attention from researchers on twitter is drawn because of valuable user's opinions.

In [15], though data used for stock market prediction is effective, data set which is used for train the model is very less. Regression algorithms efficiency can be enhanced by standardizing the data which is shown in [16], where all the real values are converted into ordinal values. In [17], new approach based on logistic regression model predicted the stock price trend of next month by using the stock prices of current month. In this approach they have not considered the social media data which would have increased the accuracy. In [18], we can see the impact of news from different web sites for the future market price by considering two market places i.e. America and Chinese.

III. Proposed Work

This paper proposes a big data model for stock market prediction using social media analysis and machine learning algorithms.

A. *Social media analytics for prediction*

The process of social media analytics used follows three step process as shown in Fig. 2.

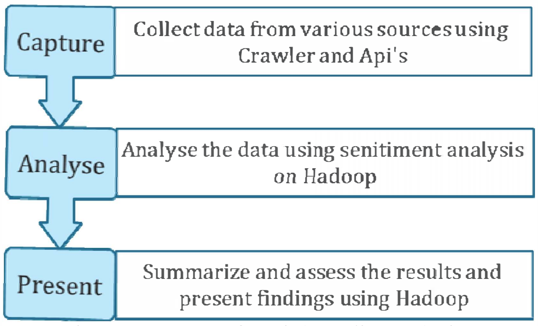
Capture
Collect data from various sources using Crawler and Api's
Analyse
Analyse the data using senitiment analysis on Hadoop
Present
Summarize and assess the results and present findings using Hadoop

Fig. 2. Process of social media analytics

For stock market, trend of the company in news and social media plays an important role on price of the stock. Predicting the sentiment of these news articles and social media data helps in predicting the rise or fall of stock price [6][7]. Hence sentiment analysis technique is used for prediction [9].

For predicting the sentiment of the stock news data, social media data of different companies over a week is considered. Model is designed to predict the sentiment for the next day. Data is collected and analyzed for four weeks and compared. Proposed prediction using sentiment analysis provides three indicators positive, negative and neutral. The model comprises of five modules as Data gathering, Data preparation, Sentiment analysis, Aggregation, and Visualization.

1) ***Data gathering****:* News articles of different companies are collected by using Mozenda Web Crawler. Tweets of these companies are collected using Twitter search API's. However for real time analytics using big data as when the data is generated it can be streamed on to HDFS using flume, storm. In this paper the results of two companies are shown. Data collected for company I comprised 355 news articles and 430 tweets for a given set of keywords of company1, for one week. Company II had 510 news articles and 599 tweets. The small data set is considered for experimental purpose. However the model is applicable for real time data analysis as well.

2) ***Data preparation***: News articles and Tweets collected are prepared for analysis by applying the following steps
   a) Lemmatization: is the process of reducing different modulated forms to a common base form so they can be analyzed as a single item.
   b) Removing Stop words: Extremely common words which are of little value to the sentiment of the news or tweet are considered as stop words and removed.
   c) Removing URL's: URLs are commonly seen in tweets but are of little help in analyzing the sentiment of the tweet, hence are removed from the tweet and news
   d) Removing duplicates: Redundancy is common in tweets because of retweet option. This will hinder the actual sentiment of the company. Hence duplicate news article and tweets are removed from the set.

3) ***Sentiment analysis***: Cleansed data from step two is put in HDFS for sentiment analysis. Hive script is run for getting sentiment of the news and tweets using the following algorithm

```
Algorithm:
For each row
Sentiment [row] =0
    For each word in the row
        Compare the word in the domain specific
        dictionary and apply sentiment sent_word;
        Sentiment [row] += sent_word;
```

4) ***Aggregation***: The sentiments of the tweets and news articles are aggregated to give the sentiment of the company.

5) ***Visualization***: The results obtained are plotted using Rhadoop. It can be observed from the graph that the sentiments reflected the trend in the stock market fluctuations. It can be observed from Fig. 3 and Fig. 4 that the sentiment prediction at the end of the week 1 and week 2 for company I predicted correctly. However for Fig. 5 and Fig. 6 show that company II prediction for week1 was not correct , but for week 2 it correctly depicted the stock market trend.

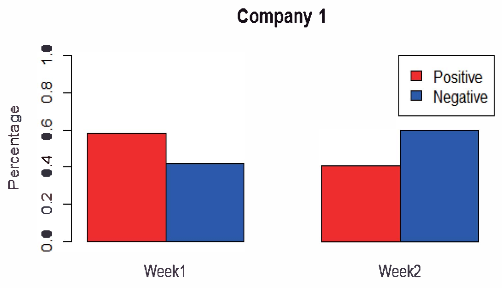

Fig. 3. : Sentiment predictions of company I two weeks

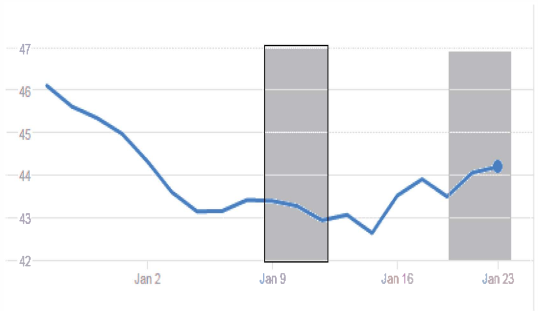

Fig. 4: Sentiment comparison with the stock market price of Company I

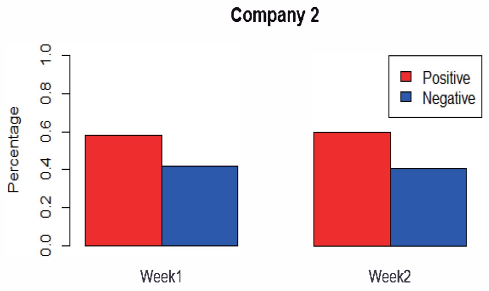

Fig. 5. : Sentiment predictions of company II two weeks

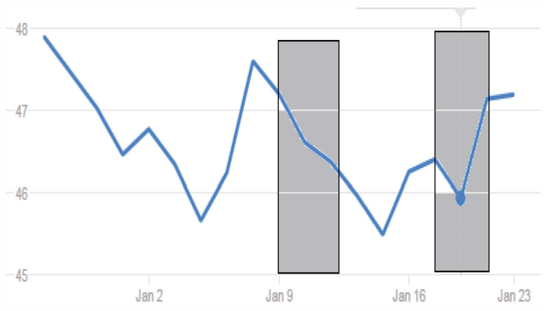

Fig. 6: Sentiment comparison with the stock market price of Company II

B. Stock market prediction using Machine Learning

Logistic regression is a machine learning algorithm for statistical classification method where an outcome is determined by one or more independent variables. Nature of outcome is categorical. Logistic regression model can be represented as

$$p = \frac{1}{1 + e^{-z}}$$

Where $z = \log\left(\frac{p}{1-p}\right)$ is a logit function.

p is probability which indicates trend of stock market.

Assume our model consists of *m*-vectors in d-dimensional feature space. For a point *x* in feature space, project it onto *m* to convert it into a real number *z* in the range -∞ to + ∞

$$z = c + m.x = c + m_1x_1 + m_2x_2 + \ldots + m_dx_d$$

Stock market is a complicated system. Forecasting of stock market trend prediction is characterized by hidden relationships, noise, data intensity, and high degree of uncertainty. In stock market trend forecast, the trend of next day's price 'up' or 'down'. So for the stock market trend prediction, logistic regression model suits well.

Data source:Yahoo Finance [19]
Data Used: 3980 rows
Testing: 50% of data i.e. 1990 rows
Training: Rest data, i.e. 1990 rows
Tool used: R Hadoop

The data can be collected according our requirement i.e. daily, monthly or yearly. For our work we have collected Daily stock market data. The data collected from Yahoo finance has many features like open price of stock (Open), close price of stock (Close), Highest price of stock on that day(Highest), Lowest price of stock on that day(Lowest). When correlation is compared between all these variables, it found to be above 0.9 for pairs. Other features like Volume traded (Volume), Adjustment Close (Adj.Close) are having very less correlation with other variables. Features which are highly correlated are considered for analysis.

A new column is added to the dataset which contains binary values '0' and '1'. For each row of dataset binary value is assigned based on the following formula.

$$trend_i \begin{cases} 0 \; if \; (Close_i - Close_{i-1}) < 0 \\ 1 \; if \; (Close_i - Close_{i-1}) > 0 \end{cases}$$

Where *i* is any row i.e. it indicates a day.
$Close_i$ Is close price of stock on day *i*.

Highly correlated features are trained for logistic regression. Model is built from training data. Same model is tested on testing dataset. Since all the features were highly correlated we have got around 70% accuracy which can be calculated from the confusion matrix.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Where TP is true positive, trend is up and it is been classified as up.
TN is true negative, trend is down and it is classified as down.
FP is false positive, trend is down but it is classified as up.
FN is false negative, trend is up but it is classified as down.

i. Model used for prdiction

Data is collected from yahoo finance. 50% of the collected data is considered for training dataset, rest 50% is considered for testing.

Generalized linear model with family binomial can be used as logistic regression model. glm function which is available in R is built under caret package. General form of glm is as shown,

glm(formula, family, data)

Formula could be any expression in the form of y ~ x , where y is binary variable, and x is list of training parameters. For example, trend ~ (Open, Close, High, Low, Adj. Close). Accuracy using this function is 70%.

The below graph Fig. 7 also depicts the 7 days observations versus predictions from the model but for the another company. The red color line indicates the company's actual trend for the next day and the line in green color shows the company's predicted trend for the next day based on the model.

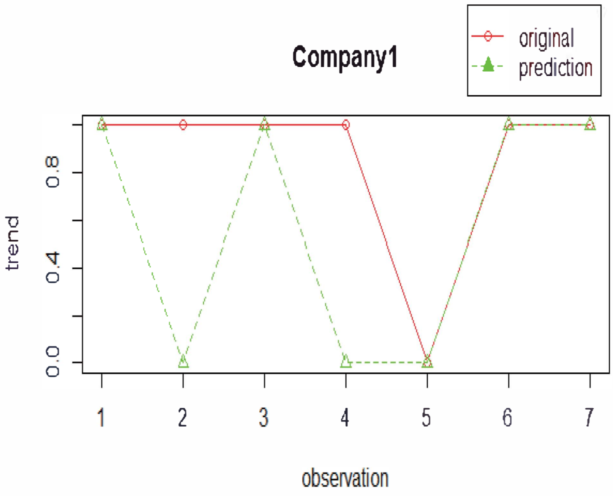

Fig. 7 Predicted data versus observation for company 1

The below graph Fig. 8 also depicts the 7 days observations versus predictions from the model but for the another company. The red color line indicates the company's actual trend for the next day and the line in green color shows the company's predicted trend for the next day based on the model.

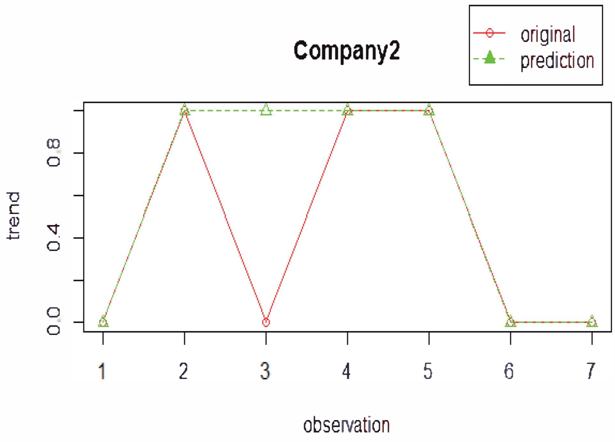

Fig. 8 Predicted data versus observation for company 2

IV. Conclusion

In this paper techniques and developed facilities for exploiting especially financial news, social media data and analysis results are presented. A prediction model has been built that uses big data analytical capabilities, social media analytics and machine learning to periodically predict the trend about stock markets. Model shows that sentiment analysis of the social data complements proven technical analysis methods such as regression analysis. It shows that volatility of the markets and the future performance of the system is affected by the economic and political news and influence of the social media.

Exploiting social media data in addition to numeric data increases the quality of the input and gives improved predictions. The aide of big data technology allows predictions at real-time.

However the algorithm used for sentiment analysis uses summative assessment of the sentiments in a particular news article or tweet, this could be improved for better sentiment calculations, which would improve the accuracy of the prediction.

References

[1] Hellstrom T. A Random Walk through the Stock Market, Licentiate Thesis,Department of Computing Science, Umea University, Sweden, 1998

[2] Lawrence Shepp, "A Model for Stock Price Fluctuations Based on Information", IEEE Transactions On Information Theory, Vol. 48, No. 6, June 2002

[3] Burton G. Malkiel, "The Efficient Market Hypothesis and Its Critics", *The Journal of Economic Perspectives,*Vol. 17, No. 1 (Winter, 2003), pp. 59-82

[4] *László Gerencsér, Balázs Torma and Zsanett Orlovits*, "Fundamental Modelling of Financial Markets," ERCIM News ,vol 78, pp 52,Jul. 2009

[5] R. Chodhury, K. Garg, "A hybrid machine learning system for stock market forecasting", Proceeding of World Academy of Science, Engineering and Technology, vol. 29 (2008) ISSN 1307-6884

[6] Arthur J. O'Connor. 2013. The Power of Popularity: An Empirical Study of the Relationship between Social Media Fan Counts and Brand Company Stock Prices. *Soc. Sci. Comput. Rev.* 31, 2 (April 2013), 229-235.DOI=10.1177/0894439312448037 http://dx.doi.org/10.1177/0894439312448037

[7] Tan, W., M. B., Blake, I., Saleh, S., Dustdar. 2013. Social-Network-Sourced Big Data Analytics. IEEE Internet Computing , 17(5):62-69

[8] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. Journal of Computational Science, 2(1):1–8, 2011

[9] Wuthrich, B.; Cho, V.; Leung, S.; Permunetilleke, D.; Sankaran, K.; Zhang, J., "Daily stock market forecast from textual web data," *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on* , vol.3, no., pp.2720,2725 vol.3, 11-14 Oct 1998

[10] S.Kopperundevi and DR.A.Iyemperumal., Stock Price Prediction of Oil and Gas Corporation using Modified Genetic Algorithm Simulated Annealing Approach. ***Aust. J. Basic & Appl. Sci.,*** *8(9): 375-382, 2014*

[11] Mizumoto K. et all.(2012) Sentiment Analysis of Stock Market News with Semi-supervised Learning. Presented at 2012 IEEE / ACIS 11th International Conference on Computer and Information Science

[12] Banic L. et al.," Using Big data and sentiment analysis in product evaluation", Presented at MIPRO 2013, Croatia, May 2013.

[13] Bing L. et al, " Public sentiment analysis in twitter data for prediction of a company's stock price movement.", Presented at IEEE 11th International Conference on e-Business Engineering,2014

[14] J. Leskovec, L. Adamic and B. Huberman. " The dynamics of viral Marketing" . In Proceedings of the 7th ACM Conference on Electronic Commerce. 2006.

[15] Zhen Hu. et all(2013) " Stock Market prediction using Support Vector Machines"., 6th International Conference on Information Management, Innovation Management and Industrial Engineering, 2013

[16] H. L. Siew and M. J. Nordin, "Regression techniques for the prediction of stock price trend," in Statistics in Science, Business, and Engineering (ICSSBE), International Conference on Langkawi: Universiti Kuala Lumpur, pp. 1-5, 2012

[17] Gong J. and Sun S.," A new approach of stock price trend prediction based on logistic regression mode", International conference on New Trends in Information and Service Science, 2009

[18] Lin N. et all." How Web News Media Impact Futures Market Price Linkage?", Sixth International conference on Business Intelligence and Financial Engineering, 2013

[19] Online: finance.yahoo.com accessed on January, 2015