A Survey on Stock Prediction with Statistical and Social Media Analytics

Indumathi S[1], Shreekant Jere[2]

[1]*M.tech, Computer Science and Engineering, REVA ITM, Bangalore, India*

[2]*REVA UNIVERSITY, Bangalore, India*

---***---

Abstract - *Stock prediction has recently grown to be a huge research area in the field of predictive analytics, big data analytics and statistical analysis. The field of stock prediction has used machine learning techniques as well as recently predictive analytics to predict stock prices. This paper gives a brief description of big data analytics and stock prediction and its methodologies. We go on to describe the current methods of stock prediction methodologies that exist in the current system where the prediction methodologies follow statistical analysis methods and predictive analysis over social media and hybrid of the two as well. On describing the current approaches, we show that there is a lot of improvement and research going on in the field of stock prediction and it will continue to improve the precision of stock prediction in the future with more research going on.*

Key Words: **Machine Learning, Predictive analysis, sentiment analysis, Natural Language Processing.**

1. INTRODUCTION

Big data analytics is the method of analyzing large amounts of data to find patterns or correlations within the data. The kinds of data that are analyzed are not necessarily structured in their formats. The unstructured data is in the form of text and not in the formats of rows and columns, which makes it challenging to understand the text and mine the particular data required within the text. Hence, big data analytics requires the processes of data mining which could be from various sources such as historical data or real time social media data. Hence forecasting, optimization, text analytics and predictive analytics play a major role in big data analytics.

Data mining is the method to compute patterns and correlations within a large amount of data. This may involve methods such as machine learning, artificial intelligence and statistical analysis. Data mining has been used to find patterns such as shooting percentage of a basketball player or to optimize the lineup of the team [1].

Predictive analysis is the process of collection of huge amounts of data to find the underlying trends and patterns in the data to make scientific decisions in the future. There have been recent advancements in stock prediction by predictive analytics of social media [1, 2] and at the same time it has been shown that machine learning techniques can be successfully used to predict stock prices [3].

Stock prediction is a key research interest in recent times and there have been many improvements in the prediction methodologies of stock. Stocks represent the claim in a company's assets. Stock prediction is the science of determining the future value of a stock. The basic methodologies of stock prediction are classified into fundamental analysis and technical analysis. Fundamental analysis does not take into account the previous time series data and rather takes into account the intrinsic worth of the stock such as the earning potential, the portfolio of the company and it is in other words like a survey an analyst would do. Technical analysis on the other hand uses previous data of stock, which can be taken on a daily, weekly or monthly basis, and this time series data is analyzed to predict the future price of the stock. The random walk hypothesis on the other hand states that past data cannot be used to forecast the future data because it simply is independent of each other although it has been successfully predicted in the recent times, which our paper explores.

The time series data can be analyzed in methods such as Auto Regression Moving Average (ARMA) and Auto Regression Integrated Moving Average (ARIMA). Another architecture used is Artificial Neural Networks (ANN). ANN has an input layer and many hidden layers may be present between the input and the output layer where the final hidden layer presents the output to the output layer. The neural network is often trained repeatedly many times until the output of the network correlates with the target output required.

2. RELATED WORK

This section of the paper contains a brief summary of the existing methods of stock prediction performed with different algorithms and approaches.

Farhad Soleimanian et al [4] have used a linear regression method to predict stock prices. Regression method is used to predict a numerical value by obtaining the past values on either daily or monthly basis, which consists of the parameters such as stock value, the opening price, closing price, lowest and highest price of the day along with the adjustments if any. To perform this, they calculated the correlation between the independent variables. By the comparison of their results and the actual stock values, they obtained a similarity of 61.35%.

Robert P. Schumaker et al [5] have used a combination of financial news as well as stock price quotes to predict the stock prices. The financial news articles, which are in text format are analyzed for the specific keywords. They specified that the bag of words approach is the easiest to implement and at the same time, it was found to be the least effective [6]. They have obtained the proper nouns from the text in news articles to perform the analysis. With the text and the stock quotes, they built a machine-learning algorithm with support vector regression. They predict the stock price after 20 minutes. They gathered about 9,211 news articles. They only took into considerations the articles that were published during the time when the stock market was open. They obtained a result of 8.5 % return on trades. They point out that their successful prediction was mainly due to the analysis financial news article analysis. Hence, the research in the direction of financial text mining has proved to be promising.

Wanjawa et al [7] have used artificial neural networks to forecast stock prices. Artificial intelligence method uses learning agents where the agent learns the patterns of the past events. In unsupervised learning, the agent learns these past patterns without feedback of any person where as in supervised learning; the agent is given the data with appropriate inputs and outputs to learn from. Artificial neural network is an example of such artificial intelligence based learning models. They proved that it is possible to predict stock using artificial intelligence based learning agents.

Stefan Nann et al [8] have used predictive analytics to predict stock prices based on social media data of Twitter and data from Yahoo! Finance as well to predict daily stock prices and hence limit or eliminate the stockbroker fee as well as the transaction costs. They obtained nearly 290,000 messages related to stocks S&P 500 index over a six-month period. The first step to obtain these messages using Twitter was to use the cash tag "$" which is the stock ticker symbol. For example, $MANGCHEFER is the tag used for obtaining tweets related to the company Mangalore Chemicals and Fertilizers. They have also obtained messages from Yahoo! Finance message boards related to the company. They have applied sentiment analysis to these forum posts, Twitter data and traditional news using Naïve Bayes classifier with a bag of words approach and POS tagging along with spam filtering built on keywords. They used the first few hundred tweets to train the model and hence add the words such as "buy", "long", "call" which provide the positive sentiment whereas words such as "sell", "short", give out the negative sentiment towards the stock. Based on their results, nearly 60 percent of the sentiments obtained for the stocks were predicted accurately. They considered over 800 virtual trades and attained a positive return of investment of up to 0.49 percent.

Ayodele et al [9] have used fundamental analysis as well as technical analysis to predict the stock prices and hence, created a hybrid approach that combines both fundamental and technical analysis. They identified 18 input variables to perform the analysis using artificial neural networks with multilayer perceptron model. The input variables of technical analysis included opening price, closing price, day high and day low price whereas the variables input to the fundamental analysis included the rumor to buy or sell the financial status of the company and so on. The hybrid approach results was found out to be an improved one compared to the result of just the technical analysis and the predictions were found to be adequate to be used as a guide for the investors.

Jianfeng et al [10] have proposed a new approach of semantic stock network (SSN) where the network nodes are the companies and the edges represent the correlation between the companies. For example, $aapl (Apple) is related to $goog (Google) with a strong correlation which is also specified in Tweets where they specify that $aapl is losing customers as everyone is buying Android phones in $goog. They have used the data from Twitter to create the SSN, which is a financial network on stock. The stock network created based on co-occurrences of tweets with ticker symbols by the results provided show a good improvement to predict stock based on sentiment analysis. Hence, using the semantic stock network along with tweets obtained from neighbors with strong correlations has shown significant improvement to predict stock.

Tina Ding et al [11] have used time series data as well as sentiments obtained from Twitter data to predict the stock prices. The sentiment analysis is performed using NLTK which is an open source suite based on Python. It also has a Naïve Bayes classifier with inbuilt training methods. They imported the data from Yahoo! Finance. They also obtained tweets regarding financial data with for specific keywords such as "sell". They trained the model with support vector machine, logistic regression and artificial neural networks and found out that the support vector machine outperforms the other two in terms of accuracy to predict stock. The results with combination of both Twitter data as well as time series data yielded the same result where the support vector machine provided the better results out of the three methods. The sentiment analysis method they used was based on just the keywords without the analysis of the context of the entire tweet. They also suggested that tools that are more sophisticated could be used to perform sentiment analysis with better accuracy.

Ganesh Bonde et al [12] have performed stock prediction by evolutionary methods of evolution strategies and genetic algorithms. Genetic algorithm is a search algorithm based upon biological evolution, which looks at fitness of the offspring i.e. the accuracy of the results. They used separate datasets for training the algorithm as well as testing it. The highest accuracy obtained from usage of genetic algorithm was 73.87%. Looking at the fact that the results have been over 70 % for every case, they suggest that there is room for improvement in the evolutionary methods used to improve the accuracy.

Xiao Ding et al [13] proposed a deep learning approach for event driven prediction of stock. The events in this case are taken out of financial news in the form of text. The bag of words approach does not capture the entire meaning of the sentence at times. For example, "Company A sues Company B" when analyzed word by word is not as effective as analyzing the sentence by its subject and the object such as

Company A being the subject and Company B being the object and the action performed is suing. This is performed by semantic analysis of the sentence by deep learning method. They specify that predicting stock prediction on a daily basis has proven to be more accurate compared to predicting stock on a weekly or a monthly basis [14, 15 and 16]. They have used historical news data treated as event sequences and performed semantic analysis over the sentences with a convolution neural network (CNN). The market simulations show that their model provides higher profit compared to the previous methods. Their results show that deep learning methods used with financial news obtained from Reuters and Bloomberg have shown with simulations a return of net profit of $16,785 with the investment of $10,000.

Xiaotian Jin et al [17] state that the stock market is a major part of the country's economic development and the capital market of the country as well. They have predicted the stock price by three statistical methods i.e. SVM regression model, least regression model and ridge regression. SVM and least regression models can have nonlinear functions whereas ridge regression model has a co-linear function for the estimation of the stock price. These methods help to obtain the independent variable that governs the stock prices based on the historical data. They used N-gram algorithm to analyze the sentiments of the people from the social media data obtained from Twitter. They have used LingPipe [18], which is open source software for natural language processing toolkit. They used a bullishness index [19] to specify the emotional index along with distinguished sentiment index [20]. They note that there is a strong correlation between the results obtained in analysis of tweets to the actual stock price of the dates.

3. CONCLUSIONS

For stock prediction, the time series data are readily available with years of daily information on stock. A lot of Twitter data as well as financial news articles are mined for information relating to the prediction of stock prices. We have seen many algorithms developed for the stock prediction in terms of technical analysis that includes support vector machines, artificial neural networks and logistic regression. At the same time, there have been many improvements to mine the data and analyze it accurately obtained from textual sources such as social media sites like Twitter and financial news articles. It is not just that the sentiment analysis using Twitter and news articles are performed with respect to the positive and negative word list but also with respect to litigious, superfluous words as well which provides an added measure of social media analytics in the field of stock prediction. We further note that there is a lot of research still going on to increase the accuracy of stock prediction in both sentiment analysis as well as technical analysis with different methodologies used to improve the accuracy. Hence, the research aims to make it easier to predict stock by taking into account not only the technical analysis but also the social media analytics as well as fundamental analysis and hybridized approaches, which include all the three methods to provide the maximum positive returns on trading.

REFERENCES

[1] Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". Procedia-Social and Behavioral Sciences, 26, 55-62.

[2] Karabulut, Y. (2011). "Can Facebook predict stock market activity?" http://bus.miami.edu/umbfc/_common/files/papers/Karabulut.pdf

[3] Berry, M. J., & Linoff, G. S. (2004). Data mining techniques: for marketing, sales, and customer relationship management. Wiley. com.

[4] "A linear regression approach to prediction of stock market trading volume: a case study" Farhad Soleimanian Gharehchopogh, Tahmineh Haddadi Bonab, Seyyed Reza Khaze, International Journal of Managing Value and Supply Chains (IJMVSC) Vol.4, No. 3, September 2013.

[5] "A Discrete Stock Price Prediction Engine Based on Financial News", Robert P. Schumaker, Hsinchun Chen, Computer, vol.43, no.1, pp. 51-56, Januray 2010, doi: 10.1109/MC.2010.2

[6] Schumaker, R.P. and Chen, H., Textual Analysis of Stock Market Prediction Using Financial News Articles. in 12th Americas Conference on Information Systems (AMCIS-2006), (Acapulco, Mexico, 2006).

[7] "ANN Model to predict stock prices at stock exchange markets", B.W. Wanjawa, L. Muchemi.

[8] "Predictive analytics on public data – the case of stock markets", Stefan Nann, Jonas Krauss, Detlef Schoder, Proceedings of the 21st European Conference on Information Systems.

[9] "Stock Price Prediction using Neural Network with Hybridized Market Indicators", Adebiyi Ayodele A, Ayo Charles K, Adebiyi Marion O, Otokiti Sunday O, Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 1, Jan 2012.

[10] "Exploiting Social Relations and Sentiment for Stock Prediction", J Si, A Mukherjee, B Liu, SJ Pan, Q Li, H Li. EMNLP 2014.

[11] "Stock Market Prediction based on Time Series Data and Market Sentimen", Tina Ding, Vanessa Fang, Daniel Zuo.

[12] "Stock price prediction using genetic algorithms and evolution strategies", Ganesh Bonde, Khaled Rasheed, The 2012 International Conference on Genetic and Evolutionary Methods.

[13] "Predictive Analytics for Sales and Marketing", Trip Kucera, David White, January 2012, Aberdeen Group.

[14] Boyi Xie, Rebecca J. Passonneau, Leon Wu, and Germ´an G. Creamer. "Semantic frames to predict stock price movement" In Proc. of ACL, pages 873–883, 2013.

[15] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. "Using structured events to predict stock price movement: An empirical investigation" In Proc. Of EMNLP, pages 1415–1425, Doha, Qatar, October 2014. Association for Computational Linguistics.

[16] Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. "More than words: Quantifying language to measure firms' fundamentals", The Journal of Finance, 63(3):1437–1467, 2008.

[17] Xiaotian Jin, Defung Guo, Hongjian Liu, "Enhanced Stock Prediction using Social Network and Statistical Model" 2014 IEEE Workshop on Advanced Research and

Technology in Industry Applications, September 2014, pp 1199-1203.

[18] http://alias-i.com/lingpipe/

[19] Oh C, Sheng O. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement[J]. 2011.

[20] Baker M, Wurgler J. Investor sentiment in the stock market[J]. 2007.