SURVEY OF STOCK PRICE PREDICTION USING SENTIMENT ANALYSIS

ANKIT SINHA[1], YASH AGRAWAL[2], VIDHAN KUMAR[3], CHANDAN KUMAR[4]

[1-4]*(Department of Computer Science & Engineering, Dayananda Sagar College of Engineering)*

[5]Assistant Professor POORNIMA KS

Assistant Professor, Computer Science Department, Dayananda Sagar College of Engineering

---***---

Abstract: *The forecasting of the stock market price is very important in the planning of business activities. The prediction of the stock price movement has gained a lot of interest from the researchers across various disciplines including computer science, economics, finance, statistics, etc. Analysis of the stock market is now a multidisciplinary process involving social media analysis, using insights from news headlines, crunching past data and using technologies like Machine Learning, Deep Learning, Time Series Analysis, Sentiment Analysis, NLP, etc. In this paper, we give an overview of how the above technologies and methodologies had been put into practice and what were their major shortcomings. The paper further talks about the essence of customizing datasets to collect raw, unstructured data from various industries and not stick to only a few giants. Our regression model works based on sentiment analysis over news headlines using NLP. Furthermore, the model is tested on live data to evaluate how well the model identifies irregularities in real industrial data.*

Keywords*: Sentiment Analysis, Natural Language Processing, Time Series Analysis ,Deep Learning, Regression, Backpropagation.*

1. INTRODUCTION

In recent years, stock market analysis has been an important consideration for companies and industries to align their business strategies. A stock market is a place where sellers sell their company's stocks while the buyers aim to increase their worth by buying stocks. However, deciding which stock to buy and sell depends on how the company is expected to function in the future. It is very beneficial in market analysis and for the investors to choose the stock to invest in. Thus stock market analysis involves handling this volatile nature of stock data.Many machine learning models are being studied such as a naïve Bayesian Regression,

Artificial Neural Networks, Support Vector Machines(SVM), Random Forest, etc. Many studies compare the performance of these models. The most important part of the machine learning techniques is the dataset used for predicting the movement of stock prices. The dataset should be as concrete as possible because a little change in the data can perpetuate massive changes in the prediction of the prices. The recent advancement in the stock analysis involves using sentimental analysis methods and Natural Language Processing (NLP) to identify the sentiments involved in news headlines of financial trading. In the stock market, text data such as news articles would have a significant impact on stock prices With sentiment analysis applied to news headlines, we aim to identify the value of data points on various emotions to understand the impact of news on the general sentiments of the public. It is often debated if it is either the stock market that drives the sentiments of people or the sentiments that run the stock market. Time Series Analysis is also one of the primary tools used today by many industries for autonomous trading. Time series analysis is used to find the dependency of variables with one another based on past data. Usage of multiple time series is another extension that focuses on finding causality between two-time series simultaneously.

2. RELATED WORKS

Prediction of stock prices is a very challenging and complicated process because price movement just behaves like a random walk and time-varying. In recent years various researchers have used intelligent methods and techniques in stock market prediction for trading decisions [8].

Therefore it is not much of a surprise that experts in machine learning and pioneers of deep learning have attempted to unravel the mystery of unstable data. According to Nagar and Hashler time variation of news headlines and extracting sentiments over a long range of time period helps in identifying strong linkages concerning stock price movement. Other researchers like Deng et al. [10] focused on the aggregation of both technical and sentiment analysis to build their models. However, the limited number of attributes along with a bad mix of variable type were the prime reasons for the average performance shown by the model. Many researchers have focused on applying neural networks to analyze the stock market[5][16],[17][4]. Siekmann et al. (2001) [19], who used fuzzy logic parameters to train the neural networks. Although the system worked well to identify the hidden uncertainties behind data points but the fact that drawing inferences on such a system require high expertise cannot be overlooked. M. Billah [17] had suggested improvements on neural networks by the use of a training algorithm which they had designed on their own. The feasibility of applying two machine learning models, Support Vector Machines (SVM) and Back Propagation Neural Network (BPN), to financial time-series forecasting for the futures trading in Indian derivative markets, was checked and it was found that SVM forecast better than the BP algorithm[3].

In the stock analysis, SVM is the most sought model owing to its predictive power, especially when using a strategy of updating the model intermittently[3].

Whenever people read newspaper articles, it is inherent that they will assign a certain emotion with the article. Analyzing human emotions on a certain piece of information is not something pristine. The WordNet Affect Lexicon [20] did a manual assignment of a word and its synonyms a raw score to place in categories. This is a very raw and brute force method of assigning scores to emotions as manual scoring cannot be applied to a large set of headlines. Shangkun DENG[10] in his work examined the results of sentiment analysis on the comments made by the general public. However, the contextual dynamics (i.e., word order) of news and comments were not considered. Applying more deep text mining for sentiment analysis of news and comments could have resulted in a better model efficiency.

Since sentiment analysis using toolkits[12] are not able to successfully find data trends, usage of NLP along with sentiment analysis have gained prominence. The various steps like Stemming, Bag of Words, Count vectorization helps in reducing the sparsity in data. SVM minimizes the upper bound generalization error rather than minimizing the training error rate. Moreover, SVMs are robust to overfitting, eventually resulting in better generalization performance than the BP neural network [4]. Kalyani Joshi[9] advanced one step above in the stock analysis game when she used a polarity detection algorithm labeling news and making the training set using a dictionary-based approach. It was found that even Random Forest performed better than SVM.

Time-series analysis is a basic concept within the field of statistical learning that allows the user to find insights from data collected over time. Jiayi Yao,[7] in his work on stock analysis on real stream data. However, it is an observed pattern that the data trends are found only for a particular time period while for another period there may not be any such pattern. Another concept widely used is Granger causality [1] Generally Granger causality is a highly effective method since it uses multi-time series. An obvious downside of granger causality is that it fails to identify any third time series which can affect the data pattern.

Our approach differs from others that we will compose data from various industries so that we get a mix of data. The model built will target the performance of companies in various stock exchanges like NASDAQ, Bombay Stock Exchange, etc. We will also use various techniques mentioned by K. V. Sujatha and S. M.Sundaram[18] which may often arise during the working of the system on handling abnormal circumstances and cause disruptions or lead to inaccurate predictions

3. MODEL DESCRIPTION

The target variable in our problem experiment is Final closing price of stock for each company's assets. Since the closing price is a quantitative variable, the problem turns out to be a regression analysis. The process from model development features various steps in succession ranging from data collection, data pre-processing, exploratory analysis, Regressor construction, Hyper-parameter optimization and model evaluation. Another factor involves feature extraction. Feature extraction is the process of identifying the attributes which have a high correlation with data and removing the ones having low linkages. The later stage of the project deals with assessing model stability, dimensionality reduction and implementing various curve metrics like ROC curve to evaluate our model's performance on stream data.

3.1 DATA COLLECTION

Generally from a trading perspective companies require mixed data ranging over a longer period of time. In the stock market, it is often a scenario that analysts can find a general trend for a portion of data, say for the first 5 years however, fail to identify correlation and linkages for the external set of data. Thus we plan to club data from various companies. Another crucial aspect when it comes to data collection for stock analysis is variance. Data pertaining to one field is somewhat bound to behave in a similar way for all the companies in that field. For instance, ridesharing companies like Uber, Ola, Lyft will find that their stocks maintain a standard curve even though the company witnesses a loss compared to healthcare companies who may tend to go to bankruptcy if their stocks fall. Thus another crucial aspect of data collection is to collect data from variegated sectors like healthcare, electronics, technology, retail, e-commerce, etc. We will build a customized dataset where we use the data released by the companies in these sectors over a time period of 10 years. The data will be scraped from Yahoo Finance Datasets, NASDAQ and specific webpages of each company. The next step involves collecting news headlines from various sources since our model will be trained on news headlines and text using NLP and sentiment analysis[6][13]. The news will be scraped from websites like The Economist , Forbes, Business Today, etc. The headlines will feature top headlines for each day corresponding to the keyword of companies. Sentiment analysis will be used to identify high-intensity words in a news headlines and find hidden linkages according to that.

3.2 FACTORS AFFECTING MODEL

Stock market prediction is not a straightforward procedure as stock behavior doesn't depend solely on one factor but is determined by a mix of factors. There is no direct cause that one factor enforces on another. For example, economic growth indirectly contributes to earnings growth. Thus considering these indirect factors becomes complex. Historically, inflation is found to have an inverse correlation with stock value. , on the other hand, deflation is generally bad for stocks because it signifies a loss in pricing power for companies. The dominance of complete market overstock value rather than the performance of a specific company in that market has been an issue. A sudden negative outlook for domain often hurts other well-performing stocks as people's emotions are swayed by the general notion. One of the common misconceptions for data patterns is incidental transactions, wherein a company buys stocks of a certain kind to fulfill some other objectives like filling up a hedge fund but analysts think that the purchase was based on the stocks performing well in the market. Demographics play a big role as middle-aged and young aged investors behave differently and the experience factor kicks in. Sometimes a trending stock may not behave as it should and this is called reverting the mean. So knowing that a stock is trendy may not be helpful sometimes. The political scenario, negotiations between countries or companies, product breakthroughs, mergers and acquisitions, and inadvertent events can impact stocks. For example, an investment policy introduction by The United States President can affect the trading strategies employed all over the world and hence require more weight to be assigned to such headlines. Thus accommodating these factors is an essential task that distinguishes our model from the traditional models.

3.3 MODEL DEVELOPMENT AND EVALUATION

Our main aim is to build a machine learning model based on NLP[12] and sentiment analysis on news headline obtained from various sources as mentioned in Data. Our customized dataset involves various types of independent variables like the company's opening value of the stock, high low, volume(number of trades), closing value, etc and the output label is closing weighted value of the stock for companies of various industries. Various derived variables will be created like company location, GDP, inflation, political scenario, policies, elections, industry type, revenue, company financial scene, positive or negative news about the company, business age, etc. A regression model will be built that predicts the stock prices and suggests if stocks have increased, decreased or stayed neutral over a given interval. Since the results will be pertaining to a time period, can be used to make inferences as to which stocks will be useful to invest in. One of the challenging tasks in predicting stock prices is working on real-time data. The ream of data needs to be processed to obtain useful formation from it. The model will be trained on the customized dataset but will be tested on live data of specific company. The testing will occur on everyday data released by companies and the performance of the model will be evaluated at the end of the month when the model results are compared with the actual stock published on stock exchanges.

4. FUTURE WORKS

The real challenge comes with testing the model on the live data released by companies. Some days our model is expected to perform great but on other days it may fail miserably. Hence primary scope is to update our model to keep it consistent with ever-changing data. Moreover, recently many advancements in stock detection have been made with twitter sentiment analysis[11,14] and social media analysis. If our model seems to work well with the data of news headlines, this model can be further improved to derive more insights from social networks

5. CONCLUSIONS

Stock market analysis is one of the most vexing tasks and even the top investing firms are still working to develop a self-sufficient model that can falsify the notion that machine learning models cannot outsmart the stock market.

Finding a future trend for a stock is a crucial task because stock trends depend upon several factors. In this paper, we assumed that news articles and stock prices are related to each other and news may have the capacity to fluctuate stock trends.

A major task will be to check to what extent does the assumption holds. Since the internal working of the algorithms depends on finding causality among variables, it will be important to know if there is any unnoticed causality which the model missed.

In our experiment we are dealing with data of many sectors, hence it will be important to train our model to an extent that it gives a satisfactory response in all sectors and not in one or two. The experiment involves intermediary steps like data visualization, exploratory data analysis, model development, K Folds clustering to assess model stability, Dimensionality Reduction, Hyper-parameter optimization and evaluation on many metrics like R Squared, Precision-Recall Curve, RMSE, etc. Since the stock markets constantly evolve, the experiment is open to a future update

6. REFERENCES

1. Andrius Mudinas, Dell Zhang, MarLevene.2019. Market Trend Prediction using Sentiment Analysis: Lessons L Learned and Paths Forward.(2019),1903.05440v1

2. Saif M. Mohammad and Peter D. Turney. 2013. 25th Signal Processing and Communications Crowdsourcing a Word-Emotion Association Lexicon. Applications Conference (SIU), Antalya, 2017, pp. 1-4. 29, 3 (2013), 436–465.

3. Shom Prasad Das and Sudarsan Padhy. 2012. Support prediction using an improved training algorithm of vector machines for prediction of futures prices in neural network," 2016 2nd International Conference on Indian stock market. International Journal of Computer Electrical, Computer & Telecommunication Applications 41, 3 (2012). Engineering (ICECTE), Rajshahi, 2016, pp. 1-4.

4. Li-Juan Cao and Francis Eng Hock Tay. 2003. Support [18]Siekmann, S., Kruse, R., & Gebhardt, J. (2001) vector machine with adaptive parameters in

financial "Information fusion in the context of stock index time series forecasting. IEEE Transactions on neural prediction". International Journal of Intelligent system networks 14, 6 (2003), 1506–1518.

5. V.V. Ramalingam, Vaibhav Sharma.2018.PREDICT Systems, 16, 1285–1298. STOCK PRICES USING NEURAL NETWORKS [19] Stefano Baccianella, Andrea Esuli and Fabrizio WITH HISTORICAL STOCK PRICES.International Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Journal of Pure and Applied Mathematics Volume 118 Resource for Sentiment Analysis and Opinion Mining.

6. No. 22 2018, 641-644 In Proceedings of LREC-10, 7th Conference on 807 806

7. DevShah,Haruna Isah,Farhana Language Resources and Evaluation, Valletta, MT, Zulkernine.2018.Predicting the Effects of News 2010, pages 2200-2204. Sentiments on the Stock Market.2018 IEEE International Conference on Big Data (Big Data)

8. Jiayi Yao, Shuhui Kong.2008.The Application of Stream Data Time-Series Pattern Reliance Mining in Stock Market Analysis. 2008 IEEE International Conference on Service Operations and Logistics,and Informatics.(10408383)

9. ASHISH SHARMA,DINESH BHURIYA,UPENDRA SINGH.2017.Survey of Stock Market Prediction Using Machine Learning Approach.International Conference on Electronics, Communication and Aerospace Technology ICECA 2017

10. Kalyani Joshi, Prof. Bharathi H. N., Prof. Jyothi Rao.2016.STOCK TREND PREDICTION USING NEWS SENTIMENT ANALYSIS.International Journal of Computer Science & Information Technology (IJCSIT) Vol 8, No 3, June 2016 [

11. Shangkun DENG, Takashi MITSUBISHI, Kei SHIODA, Tatsuro SHIMADA, Akito SAKURAI.2011.Combining Technical Analysis with Sentiment Analysis for Stock Price Prediction.2011 Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing

12. Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. CoRR abs/1010.3003 (2010)

13. Yasir Ali Solangi, Zulfiqar Ali Solangi, Samreen Aarain, Amna Abro, Ghulam Ali Mallah, Asadullah Shah.2018.Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis.2018 IEEE 5th International Conference on Engineering Technologies & Applied Sciences, 22- 23 Nov 2018, Bangkok Thailand.

14. Spandan Ghose Chowdhury, Soham Routh , Satyajit Chakrabarti, News Analytics and Sentiment Analysis to Predict Stock Price Trends, (IJCSIT)

15. A. Mittal and A. Goel, "Stock prediction using twitter sentiment analysis,"

16. Mohit Iyer ,Ritika Mehra.2018.A Survey on Stock Market Prediction. 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) ,(2018)18792225

17. Mohit Iyer ,Ritika Mehra.2018.A Survey on Stock Market Prediction. 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) ,(2018)18792225

18. H. Gunduz, Z. Cataltepe and Y. Yaslan, "Stock market direction prediction using deep neural networks," 201