Sentiment Analysis of News Articles and Twitter Data for Predicting Stock Price Movement

Yatesh Laxman Kumar Gumma
Department of Computer Science
Lakehead University
Thunder Bay, ON, Canada
ygumma@lakeheadu.ca

Dr. Jinan Fiaidhi
Department of Computer Science
Lakehead University
Thunder Bay, ON, Canada
jfiaidhi@lakeheadu.ca

***Abstract*—Stocks play a major role in assessing the company's success and future. The stock market has now become an obsession for millennials. It is a place where anyone can join without much effort and the main reason is for monetary gain. Any investor should be aware of two important aspects when to buy or sell a stock and the major factors that affect the stock price. The major factor responsible for price fluctuation is the sentiment of the investors which in turn depends on the news. The ideal scenario is to purchase a stock at a low price and then sell it at a higher price. That said, stock investments still remain a risky proposition for the uninitiated which leads to both analyst and researcher's attention. Accurate predictions can help investors take correct decisions about the selling or purchase of stocks. Many researchers from various disciplines, including computer science, statistics, economics, finance, and operations research, have expressed interest in stock price prediction. Recent studies have shown that new stories and social media discussions have an observable effect on investor attitudes toward the financial market. Predicting the exact value is a pointless effort, but one thing is certain we can predict the trend or direction in which the price of a stock will move. The project goal is to design and implement a predictive system that helps investors in making their investment decisions. Using sentiment analysis on the tweets, news articles, and also closing price values of various stocks, we seek to build a system that predicts the stock price movement of various companies. The present paper has employed two different algorithms LSTM and ARIMA for stock price prediction. The correlation between stock market movement and sentiments are correlation is analyzed using supervised machine learning principles of sentimental analysis.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Stocks play a major role in assessing the company's success and future. The stock market has now become an obsession for millennials. It is a place where anyone can join without much effort and the main reason is for monetary gain. Any investor should be aware of two important aspects when to buy or sell a stock and the major factors that affect the stock price. The major factor responsible for price fluctuation is the sentiment of the investors which in turn depends on the news. The ideal scenario is to purchase a stock at a low price and then sell it at a higher price. Significant losses can be avoided by making a well-informed stock purchase and sale decisions.

Earlier studies on the stock market are completely based on historical prices, but later studies debunked the approach as the prices are largely fluctuating. As per Efficient Market Hypothesis(EMH) news and current events decides the stock market price in a particular day. [1]. Stock price follows a random walk pattern which makes prediction models not more than 50% accurate as news and current events are unpredictable. To tackle the problem numerous approaches have been proposed.

Fundamental analysis and Technical analysis are two major techniques in the stock market, does not matter how much effort researchers put into it still remains a difficult problem. Recent advancements in machine learning motivated a lot of researchers and brought the focus to the field. Lot researchers used various machine learning algorithms have been used to analyze price patterns and predict stock prices, index changes. Because of non-linearity in stock data and market volatility, it is very difficult to extract valuable information from patterns. Several researchers, however, proposed novel approaches to predicting stock prices. For predicting the stock prices researchers used various type of algorithms such as Linear Regression, Support Vector Machine, etc..,. Limitation of this approach is accuracy. A better approach is needed which can predict high variations prices with significant accuracy. R Naya and P Braak [3] used the predictive power clustering technique for making stock price predictions

The appearance of social media increased the amount of information about public feelings. It evolved into a platform for people to express their feelings about various topics, and it has a significant impact on overall public opinion, which has piqued the interest of researchers in recent years. Twitter is a microblogging application that lets users follow the opinions of other users and comment on them in real-time. [4]. Every day more than a million users post over 140 million tweets which made Twitter corpus data valuable for researchers. Information extracted from tweets is very useful for making predictions.

Bollen et.al. [5] demonstrated that there is a high correlation between Twitter data and Dow Jones Industrial Average(DJIA). All of these studies show Twitter to be a valuable resource and a powerful tool for conducting research and making predictions. Asur and Huberman [6] predicted the box office collections for movies prior to their release base on public opinion on movies that are expressed on Twitter.

Nowadays, the number of news sources has increased dra-

matically. Due to the abundance of news, it is difficult for investors to determine the trend of stock prices. As a result, an automated system that predicts future stock prices will be beneficial to investors. An automated system that gathers financial news articles, stock data related to the companies and uses a machine learning model on those data to predict price movement.

In this paper, we use characteristics of two different algorithms LSTM and ARIMA on time series for stock price prediction. The collection of tweets and news articles are analyzed using supervised machine learning principles for a correlation between stock price movement and sentiment.

Remaining part of the paper is structed as follows all the past work related in this field is defined in Section II. Section III describes the methodology which contains Data Gathering, Data Pre-Processing, Sentimental Analysis, LSTM, ARIMA which is followed by Section IV where the results and performance of our models are discussed. Finally we concluded in Section V along with our future work plan.

II. Related Work

Bidirectional Encoder Representation (BERT) was introduced by Devlin et.al. [7], which is a new language model representation that enables unlabeled text to be pre-trained in all layers in both left and right. The model's results and new features enable low-level resource tasks to benefit from deep unidirectional architecture, which further leads to one of the important tools in natural language processing.

Kalyani et.al. [8] used news sentiment classification on financial news articles about a company was used to forecast its future stock trend. The relationship between stock data and news articles is determined by usinf dictonary based approach. By combing general and finance-specific sentiments positive and negative words are created for the dictionaries. They developed different classification models among which Random Forest(RF) and Support Vector Machine (SVM) are the most suitable models based on the results.

Yauheniya et.al. [9] categorized news articles based on there realtionship with stock. To learn from these categories different kernel types are used. Among all the methods used Math Kernel Library (MKL) is suitable with good results.

Pegah [10] labeled news articles with a different approach and compared them with news articles that are labeled randomly. Both datasets were passed to Support Vector Machine (SVM) where her approach has an accuracy of 83% and random labeling approach has 53% accuracy

Gaurav Jariwala .et.al. [11] compared different algorithms under the same circumstances and concluding the best model based on accuracy. The algorithms they have used are K-Mean, Naive Bayes, and Support Vector Machine.

Wang et.al. [12] in order to make appropriate decisions during the outbreak, they used public sentiment analysis to have insightful information. They used popular Chinese social media site Sina Weibo posts, which adopted an unsupervised BERT model to classify sentiment categorizes and TF-IDF model to summarize the topics of the posts.

Li et al. [13] collected the stock exchange data for five years along with financial news articles over the same period of time, to draw a correlation between stock market trends and news articles. They collected the open, high, close, and low price of stock for each company in a particular trading day.

Alostad and Davulcu [14] used hourly stock price of 30 stocks, along with news articles from NASDQ website. All the tweets related to those 30 stocks were collected for 6 months. All the articles have been processed in different ways for feature extraction. Alostad and Davulcu [14] used N-gram to extract features after the removal of stop words, white space, punctuations, and numbers. They arranged the final features into a document matrix and extracted sentences from each document using OpenNLP. To detect the sentiment they used the SentiStrength library with Loughran and McDonald Financial Sentiment Dictionaries.

For text pre-processing various papers followed these approaches removal of HTML tags [15], tokenization of sentences [15], noun phrasing [16], document weighting [15], TFIDF [15], and extraction of named entities. Alostad and Davulcu [14] solve the price trend prediction as a classification problem. They implemented logistic regression to the n-gram document matrix, stock price direction for each hour, and document weight. Later SVM is used for classification. Experiments indicated that extracting document level sentiment have not improved prediction accuracy. In other works random forest, naïve Bayesian, and genetic algorithms were used to predict stock price.

Nowadays, Big Data is very powerful and popular tool. Lee [17] estimated impact of COVID on US stock market using Google data on coronavirus and Daily News Sentiment Index (DNSI). The relationship between 11 selected US stock market sector indices and COVID-19 sentiment is investigated. Stock market crisis create a rippling effect on investors decision making based on public sentiment. The results demonstrated the distinct effects of COVID-19 sentiment across industries and classified them into various correlation groups.

III. Methodology

A. Data Collection

Data set construction can be approached in many ways. We focused on both news headlines and tweets about the particular company. So, we created a data set by web-scrapping from yahoo news and collecting tweets from twitter API. Data sets are merged by human effort based on specific conditions like date. There are numerous economic portals and libraries to fully automate the process. Tweets were collected via the Twitter API and then filtered with keywords like #AirCanada. Not only will public opinion of the company have a significant impact on stock, but so will the products and services offered by the company. Tweets are extracted carefully in a way that they accurately represent the public opinion about Air Canada over a period of time.

As mentioned earlier with news headlines the main goal is to use up to date data. All the news headings extraction is fully automated by using web scrapping technique with the

help of beautifulsoup. We could extract the data from different news sources but in our case we used Yahoo News as our data source. News that is extracted based on timestamp is saved in to CSV file. Once both the tweets and news data are saved in two different separate CSV files they are merged manually based on the timestamp. The workflow of news articles and tweets from yahoo news and Twitter is shown in figure 1 Figure 1 represents graphically the step by step process of how the news articles and tweets are collected from yahoo news and twitter.

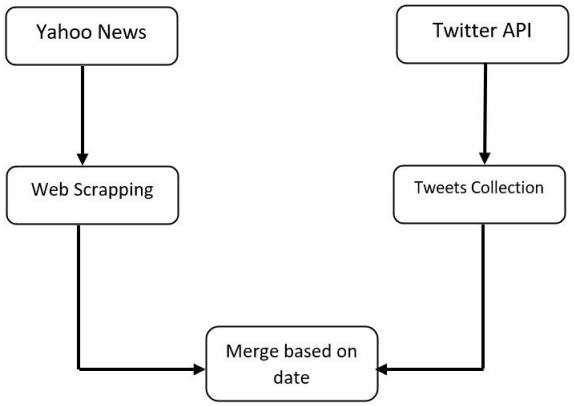

Fig. 1: News and Tweets Collection Flow Chart

Daily stock data for Air Canada is collected using Yahoo Finance. The daily features encompassed by our model include.

- Opening Price: stock price at 9.30 AM ET
- Closing Price: stock price at 4.00 PM ET
- High: Day‘s highest price
- Low: Day’s lowest price
- Volume: number of stocks traded in a day.

For binary classification tasks, we generate labels that indicate whether a stock moved up or down. If the closing price of the previous days in greater than or equal to closing price of today, we set a label of 1, otherwise it was given 0.

B. Data Pre-Processing

Tweets contain a lot of acronyms, emoticons, and unnecessary data like pictures and URLs.So tweets need to be preprocessed to represent the correct emotions of the public. Special characters from tweets are removed by using regex matching. In python, regex matching is used to match URLs and is replaced by URL terms. So all other characters other than alphabetical and number are replaced with white space. News articles that are extracted by using web-scraping will not have any unnecessary data like pictures and URLs as data is extracted based on HTML tags. In the same way as tweets, by using regex matching all the characters other than alphanumeric characters are replaced with white space. After this stage data is ready for sentiment analysis.

C. Sentiment Analysis

The Sentimental analysis task is very much field specific. Starting from human-labeled data to various deep learning methods there are many possibilities for sentimental analysis, and there are many sentiment analyzers available as open source. The main issues with these analyzers they are trained on different corpus for instance movie reviews and stock news are not the same. In this case, we compare the capabilities of TextBlob and NLTK-VADER. Main goal is to determine the sentimental value of sentence to classify as positive, negative, and neutral.

1) Text Blob: The Sentimental analysis task is very much field-specific. Starting from human-labeled data to various deep learning methods there are many possibilities for sentimental analysis, and there are many sentiment analyzers available as open source. The main issues with these analyzers they are trained on different corpus for instance movie reviews and stock news are not the same. In this case, we compare the capabilities of TextBlob and NLTK-VADER. The main goal is to determine the sentimental value of a sentence to classify as positive, negative, and neutral.

2) NLTK VADER: Natural Language Toolkit(NLTK) is one of the most powerful NLP libraries, with packages for making machines understand human language and react appropriately. With the help of SentimentIntensityAnalyzer will focus on sentimental analysis. Just like in the TextBlob polarity value of the sentence scales between -1 and 1. VADER is a SentimentIntensityAnalyzer is a component of NLTK which has four sentiment analysis scores: compound, neutral, positive, and negative. The last three scores add up to one, Input intensity is measured in terms compound score

Once the sentiment data been pre-processed, it is merged with stock data based on the date range of the stock data into a new data frame. The missing records in stock data because of public holidays and weekends, while merging dataset if sentiment data has a record on those days such records are ignored. Tweets and news articles gathered from web scraping are merged into a single column and then added to the merged data frame. The polarity and subjectivity of news and tweets are calculated using TextBlob and their sentiment scores are obtained by VADER SentimentIntensityAnalyzer. All the obtained values are appended to the data frame.Model Training:-

3) Model Training: Before feeding the data to the classification model certain columns other than the stock data, subjectivity, polarity, and sentiment scores are vomited. The label value which indicates if the stock price is increased or decreased is separated from features before they are passed to the model. The whole data is divided into testing and training with 80% in training and 20% in training.

LDA stands for Linear Discriminant Analysis [18] as the name implies reduction in the number of dimensions while preserving much information as possible. For training extracted features are fed to the LDA algorithm. Once the model is trained then the test data is passed to make the predictions.

The results of sentiment classification are discussed in the following sections.

D. Long Short Term Memory

LSTM stands for Long Short Term Memory which has the ability to predict time series data based on the number of lags. One of the advantages of LSTM are its feedback connections. LSTM model is used to predict the stock price trend, which consists of four layers input layer, hidden layer, attention layer, and output layer. LSTM architecture is useful for modeling arbitrary intervals of information.

LSTM model is constructed and trained with Keras neural network API. T The API is a Python-based open-source deep-learning library that uses Tensor-flow as a backend. The Amount of training dataset and hyperparameter settings are factors on which LSTM‘s model’s accuracy dependent. So 80% of the data is dedicated to training the model and the rest 20% is used for testing the model’s performance.

To model an LSTM problem, we predict the stock price at time t as a function of stock prices at t-1, t-2,. . . ,t-m where m is the window size of the past stock price. For instance, we consider the past 60 days of data and fed the data to model then predict the 61st day‘s value. In the next iteration, the 1st value is ignored and the 61st-day value is considered as the 60th-day value then we will predict the 62nd-day value. The iterative process is repeated until 30 days of predictions are made.

Before the LSTM model uses the dataset we collected from Yahoo Finance, we transform it to a given range using MinMaxScaler and reshape it into three dimensions. An LSTM model is built by experimenting with the window size, the number of LSTM layers, units in each layer, and dropout percentage. During the model fitting results for different batch sizes and epoch, sizes were analyzed using the ADAM optimizer during training. Model is constructed with two LSTM and two dense layers with 50 and 25 neurons in each, respectively.

The difference between desired output and LSTM model‘s output is estimated by using loss function. Overfitting and underfitting can be prevented by using loss function after each epoch in training phase. We used Mean Square Error (MSE) is one of the widely used loss function in time series forecasting.

For measuring the accuracy of model predicted values are transformed back to regular value using the inverse transformation. Both the MSE and RMSE will evaluate the LSTM model‘s. In our case we calculated the RMSE value. Both the actual and predicted value along with the RMSE value are discussed in the results section.

E. ARIMA

ARIMA is the most commonly used model for time series analysis. It consists of three parts autoregressive(AR), integrated (I), moving average(MA) models. The notation for the ARIMA model is ARIMA(p, d, q), where p, d, and q are the parameters of the AR, I, and MA models, respectively. ARIMA models are mainly suitable for short-term forecasting, which makes it one of the best models for stock price prediction.

We performed a meta parameter grid search to find the best values for parameters p, q, and d. To identify best trained ARIMA model Akaike Information Criteria(AIC) is used. AIC is widely used to measure the performance of statistical models. It quantifies goodness of the fit of a model. Models with lower AIC are better. We passed a range of (0,9) to the grid search function and finally considered the parameters which gave the lowest AIC value. We tried different models by changing the number of observations. Finally, we predicted stock prices for different durations, namely 7 days and 10 days.

We can even obtain the parameters for the ARIMA model manually. The term d can be obtained by finding out the order of differencing required to make the data stationary. Similarly p and q terms can be obtained by observing the partial auto correlation and auto correlation graphs. This approach takes lot of human efforts as every time there are change in the data terms value changes.

Both the actual and predicted value along with the RMSE value are discussed in the results section.

IV. Results and Discussion

In this section, we discuss the results obtained using our models.

A. Sentiment Analysis

In this sentimental analysis study, we have trained LDA on using two different kinds of datasets on Air Canada. The first dataset is a collection of tweets and news articles, while the other one is a collection of the top 25 news articles for the day from Reddit. Stock data of Air Canada is taken from Yahoo Finance. The accuracy of both datasets using the LDA model is shown in Table 1.

	Collection of Tweets and News	Reddit News Articles
Accuracy	87.5%	78%

TABLE I: Sentiment Analysis Accuracy

From the above results, the model is able to predict more accurately when taken both the social media sentiment and news articles into consideration compared to that of just the top 25 new articles. The articles or tweets posted from the market close to till the market open the next day have a higher predictive power in predicting market movement the following day.

B. Long Short Term Memory

LSTM model is trained with an various epoch sizes by using Air Canada closing price value. For the experiment 9 years of stock data starting from 2012-01-01 until 2021-02-16 has been extracted from yahoo finance by using panadas data reader. The performance of model is assessed by using Root Mean Square Error value which is calculated between predicted price and actual price. The model was trained with a batch size of 64, number of epochs as 1000 where we achieved good result.

Figure 2 indicates the graphical representation of both the actual price and predict price for Air Canada from April 2019 to Feburary 2021. From the graph it is clear that there is not huge difference between both values there are certain days where the difference is actually negligible. As told earlier performance of model is estimated by RMSE lower the value more accurate our model . The results indicate that our RMSE value is 1.09 which is considerably low value. Table 3 showcase the actual price and predicted price values of last five trading days.

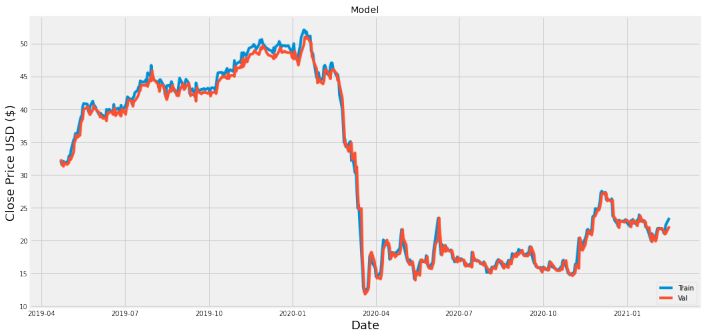

Fig. 2: LSTM Actual and Predicted Price

RMSE	1.087

TABLE II: LSTM Accuracy

Date	Predicted Value	Actual Value
2021-02-16	22.25	23.42
2021-02-12	21.13	22.33
2021-02-11	21.04	21.20
2021-02-10	21.51	21.08
2021-02-09	21.69	21.50
2021-02-08	21.94	21.65

TABLE III: LSTM Actual Price and Predicted Price

C. ARIMA

Table5 shows the predicted and actual price values for Air Canada. For the experiment 9 years of stock data starting from 2012-01-01 until 2021-02-16 has been extracted from yahoo finance by using panadas data reader. In order to evaluate the ARIMA models performance same metrics as the LSTM are used.

Figure 3 showcase the graphical representation of both the predicted and actual prices for Air Canada on ARIMA model with order (0,1,0). From the graph it is clear that there is not much difference between the predicted and actual values some instances of closely related of actual and predicted values time. RMSE value for our model is 1.01 which is less than LSTM.

RMSE	1.01

TABLE IV: ARIMA Accuracy

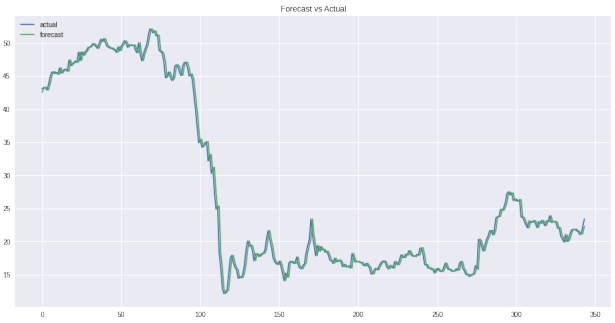

Fig. 3: LSTM Actual and Predicted Price

Date	Predicted Value	Actual Value
2021-02-16	22.33	23.42
2021-02-12	21.20	22.33
2021-02-11	21.08	21.20
2021-02-10	21.50	21.08
2021-02-09	21.65	21.50
2021-02-08	21.87	21.65

TABLE V: ARIMA Actual Price and Predicted Price

V. Conclusion and Future Work

To conclude the results show that there is a strong correlation between stock prices and investor sentiment, which is primarily influenced by news stories. It is common sense that the effect of the event or sentiment on the stock market sometimes lasts for few days, even several weeks or months. Analyzing the sentiment analysis results it is discovered that the value difference between positive sentiment score and negative sentiment score is in the range from -0.23 to 0.08. And a prediction models were built based on time series forecasting models such as ARIMA and LSTM. We achieve better results with LSTM and in the case of ARIMA if the prediction days are less. However, the LSTM model did not perform well in such cases where the stock price is fluctuating.

Future work may involve expanding the analyses and possibly adding new features. Compare stock prediction with different sentiment analysis tools. Building a domain-specific model by grouping companies according to their sector considering the adverse effect on the company‘s stock price due to news about other related companies. Moreover, considering the event or sentiment on the stock market stays longer than a day. Build a better LSTM model for predicting the price when the stock price is less or more volatile.

References

[1] E.F. Fama, The behavior of stock-market prices, The Journal of Business 38 (1) (1965) 34105, http://dx.doi.org/10.2307/2350752

[2] A. Skabar, I. Cloete, ”Neural networks, financial trading and the efficient markets hypothesis”, in ACSC 02: Proceedings of the twenty-fifth Australasian conference on computer science, Australian Computer Society, Inc., Darlinghurst, Australia, 2002, pages 241-249.

[3] R. Nayak, P.Braak, ”Temporal pattern matching for the prediction of stock prices”, in AIDM ’07:Proceedings of the second international workshop on Integrating artificial intelligence and data mining, Australian Computer Society, Inc, Darlinghurst, Australia, 2007, pages 95103

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

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

[6] Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: Proceedings of the ACM International Conference on Web Intelligence, pp. 492-499 (2010)

[7] Devlin, J., Chang, M. W.., Lee, K., & Toutanova, K (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[8] Kalyani, J., Bharathi, P., & Jyothi, P. (2016). Stock trend prediction using news sentiment analysis. arXiv preprint arXiv:1607.01958.

[9] Y. Shynkevich, T. M. McGinnity, S. Coleman and A. Belatreche, "Predicting Stock Price Movements Based on Different Categories of News Articles," 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, 2015, pp. 703-710.

[10] P. Falinouss, "Stock trend prediction using news articles a text mining approach," Dissertation, 2007, pp. 83-84.

[11] G. Jariwala, H. Agarwal and V. Jadhav, "Sentimental Analysis of News Headlines for Stock Market," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-5, doi: 10.1109/INOCON50539.2020.9298333.

[12] Wang, T., Lu, K., Chow, K. P., & Zhu, Q. (2020). COVID-19 sensing: Negative sentiment analysis on social media in China via bert model. IEEE Access, 8, 138162–138169. https://doi.org/10.1109/Access.6287639

[13] X. Li, H. Xie, Y. Song, S. Zhu, Q. Li, and F. L. Wang, "Doessummarization help stock prediction? a news impact analysis," IEEE Intelligent Systems, vol. 30, no. 3, pp. 26–34, May 2015.

[14] H. Alostad and H. Davulcu, "Directional prediction of stock prices using breaking news on twitter," in 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WIIAT), vol. 1, Dec 2015, pp. 523–530.

[15] D. Duong, T. Nguyen, and M. Dang, "Stock market prediction using financial news articles on ho chi minh stock exchange," in Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, ser. IMCOM '16. New York, NY, USA: ACM, 2016, pp. 71:1–71:6. [Online]. Available:http://doi.acm.org/10.1145/2857546.2857619.

[16] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The azfin text system," ACM Trans. Inf. Syst., vol. 27, no. 2, pp. 12:1–12:19, Mar. 2009. [Online]. Available: http://doi.acm.org/10.1145/1462198.1462204.

[17] Lee, H. S. (2020). Exploring the initial impact of COVID-19 sentiment on US stock market using big data. Sustainability, 12(16), 6648. https://doi.org/10.3390/su12166648

[18] E.F. Fama, The behavior of stock-market prices, The Journal of Business 38 (1) (1965) 34105, http://dx.doi.org/10.2307/2350752