Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2020, 8, 309-319
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks

Sai Vikram Kolasani[1], Rida Assaf[2]

[1]Trumbull High School, Trumbull, USA
[2]University of Chicago/Computer Science, Chicago, USA
Email: saikolasani100@gmail.com, rida@uchicago.edu

How to cite this paper: Kolasani, S.V. and Assaf, R. (2020) Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks. *Journal of Data Analysis and Information Processing*, **8**, 309-319.
https://doi.org/10.4236/jdaip.2020.84018

Received: October 19, 2020
Accepted: November 14, 2020
Published: November 17, 2020

Abstract

External factors, such as social media and financial news, can have widespread effects on stock price movement. For this reason, social media is considered a useful resource for precise market predictions. In this paper, we show the effectiveness of using Twitter posts to predict stock prices. We start by training various models on the Sentiment 140 Twitter data. We found that Support Vector Machines (SVM) performed best (0.83 accuracy) in the sentimental analysis, so we used it to predict the average sentiment of tweets for each day that the market was open. Next, we use the sentimental analysis of one year's data of tweets that contain the "stock market", "stocktwits", "AAPL" keywords, with the goal of predicting the corresponding stock prices of Apple Inc. (AAPL) and the US's Dow Jones Industrial Average (DJIA) index prices. Two models, Boosted Regression Trees and Multilayer Perceptron Neural Networks were used to predict the closing price difference of AAPL and DJIA prices. We show that neural networks perform substantially better than traditional models for stocks' price prediction.

Keywords

Tweets, Sentiment Analysis with Machine Learning, Support Vector Machines (SVM), Neural Networks, Stock Prediction

1. Introduction

It is in the interest of many people and companies to predict the price movement and direction of the stock market. Also, the stock market is a vital component of a country's economy. It is one of the most significant opportunities for investment by companies and investors. Stock traders need to predict trends in the stock market to determine when to sell or buy a stock. To see profits, stock trad-

ers need to acquire those stocks whose prices are expected to rise shortly and sell those stocks whose prices are expected to decline. If traders can adequately predict the stock trends and patterns, they can earn a considerable profit margin. However, stock markets are very volatile and, consequently, difficult to predict. External factors, such as social media and financial news, can have widespread effects on stock price movement. For this reason, social media is considered to have profound importance for precise market predictions.

Investors assess a company's performance and its stock before determining whether to acquire the company's shares, in order to avoid buying risky stocks. This evaluation comprises an analysis of the company's execution on social media websites. One such social media platform that has great importance in the finance and stock market realm is Twitter. One hundred million active Twitter users update nearly 500 million tweets every day [1]. Users express their opinions, decisions, feelings, and predictions through these tweets, which can be translated into useful information. However, such a tremendous amount of social media data cannot be entirely assessed by investors alone. It is a nearly impossible task for humans to perform on their own. Therefore, a computerized analysis system is necessary for investors, as this system will automatically evaluate stock trends using such large amounts of data in data sets.

A substantial amount of practice in previous research on stock prediction has been applied to historical or social media data. Research with historical data includes using a technical analysis approach in which mathematics is employed to analyze data for finding future stock market trends and prices [2]. Researchers used different machine learning techniques, such as deep learning [3] and regression analysis [4], on stock historical price data. However, these studies did not include external factors such as social media. It is important to utilize social media data because events expressed through social media can significantly affect stock prices and trends due to the belief that prices change because of human behavior which can be reflected by social media.

Social media sentiment analysis is an excellent reservoir of information and can provide insights that can indicate positive or negative views on stocks and trends. There has been a sufficient amount of research on sentiment analysis on various topics, such as movie reviews and Twitter feeds in past years. Agarwal *et al.* 2011 [5], examined sentiment analysis on Twitter Data and prefaced POS-specific prior polarity features and investigated the use of a tree kernel to eliminate the need for slow feature engineering. Pang and Lee 2004 [6], proposed an innovative machine-learning method that utilizes text-categorization techniques to just subject portions of texts. Kim 2014 [7], advised a simple one-layer Convolutional Neural Network (CNN) that would produce impressive, second to none results across several different data sets. In 4 out of the seven categories tested in the experiment, CNN did much better, whereas it was comparable to the other three types. CNN had the highest accuracy with 81.5 in movie reviews, etc. The robust results achieved with this CNN design suggest that neural networks may serve as a better replacement for well-established baseline models,

such as Support Vector Machines [8] and Logistic Regression.

Furthermore, the research that has used both social media and historical data has much room for improvement. Studies conducted on Twitter and Stock market data to predict the stock market using machine learning algorithms include Chakraborty *et al.* 2017 [9]; Khatri and Srivastava 2016 [10]; Chen and Lazer 2011 [11]; Khan *et al.* 2020 [12].

This particular research paper will build on Chakraborty *et al.*'s research paper: "Predicting stock movement using sentiment analysis of the Twitter feed". In their article, the researchers have found that Twitter data could predict stock prices very well on stable days in the stock market. However, the researchers used a boosted regression tree model to predict the stock price difference for the next day with the current day's stock market Sentiment. This paper will implement neural networks to see if they produce better results than the boosted tree model. Specifically, a Multilayer Perceptron Neural Network (MLP) model will be employed. This paper aims to improve the previous writing using MLP and analyze the effectiveness of using Twitter data to predict stock market trends and prices.

In this paper, a sentiment tagged Twitter dataset of 1.6 million tweets collected from Sentiment 140 will be used for sentiment classification. Then, the Boosted Regression Tree and Multilayer Perceptron models will be used for predicting the next day's stock movement with the present day's tweets containing the "stock market", "StockTwits", "AAPL". The hypothesis that this paper will test is: "Can Stock Market related tweets accurately predict stock market movement?" Furthermore, this paper will also test: "Are neural networks more effective at predicting the stock market movement than traditional models?"

2. Materials and Methods

2.1. Data

Similarly, to Chakraborty *et al.*, the training data set was collected through Sentiment 140 that is available on Kaggle [13]. The dataset contains 1.6 million hand-tagged tweets, collected through Sentiment 140 API. The tweets are tagged "1" and "0" for being "positive" and "negative". We perform a random split over the dataset to divide the dataset into a training dataset and a testing data set. The training dataset contains 1.52 million tweets, whereas the testing dataset contains 80,000 tweets. The distribution of the data is shown in **Table 1**.

Table 1. Data distributions.

	Training Data (# of Tweets)	Testing Data (# of Tweets)
Positive (# of Tweets with Positive Sentiment)	760,154	39,989
Negative (# of Tweets with Negative Sentiment)	759,846	40,011

As shown in Table 1, the data is reasonably balanced with an almost equal amount of Positive and Negative tweets in both the Sentiment 140 testing and training data.

Next, tweets containing the "stock market", "StockTwits", "AAPL" keywords that were posted between January and December 2016 are collected for predicting the corresponding stock movement. We assembled at most a hundred tweets every day. The tweets were collected using GetOldTweets3, which is a Python library for accessing old tweets. It allows the user to get old tweets specified by dates and keywords or usernames. It also enables the user to get tweets based on location. We use GetOldTweets3 because it allows us to access old tweets, unlike other APIs.

Stock historical price data is available on Yahoo Finance [14]. The selected stock markets' price data are collected from Yahoo Finance for the chosen period in csv file format. The downloaded data files have seven features—Date, Open, High, Low, Close, Volume, and Adjusted Close—which on a specific date, show the stock traded day, stock open price, stock maximum trading price, stock lowest trading price, stock closing price, number of shares traded, and closing price of a stock when dividends are paid to investors, respectively. In this paper, only the Date and Close price are used.

The historical stock data for this experiment was collected using Yahoo Finance. Likewise, to Chakraborty *et al.*, data was obtained on the stock close price of DJIA and APPLE Inc. from January 2016 to December 2016. Data was only collected on days where the stock market was open.

Since the stock keywords data was not tagged with sentimental numbers, the Sentiment 140 dataset models were used to predict sentimental values. Table 2 is a sample of the Sentiment 140 dataset.

2.2. Data Preprocessing

Each of the tweets will be preprocessed with the following guidelines. The preprocessing of the data will be conducted by running a function on all of the text with the following guidelines. The function will then transform the data as shown in Table 3. This preprocessing process differs from the previous study as Lemmatization, Removing Keywords, and Removing Short words were added in this research study. These steps were added as they better allow for the data to be preprocessed for sentimental analysis.

1) Lower Casing: Each text is converted to lowercase.

2) Replacing URLs: Links starting with "Http" or "https" or "www" are replaced by "URL".

3) Replacing Emojis: Replace emojis by using a pre-defined dictionary containing emojis along with their meaning. (e.g.: ":)" to "EMOJIsmile")

4) Replacing Usernames: Replace @Usernames with the word "USER". (e.g.: "@Kaggle" to "USER")

5) We are removing Non-Alphabets: Replacing characters except Digits and Alphabets with space.

Table 2. Snapshot showing the first 4 examples in the dataset.

Sentiment	Text
0	@switchfoot http://twitpic.com/2y1zl - Awww, t …
0	is upset that he can't update his Facebook by …
0	@Kenichan I dived many times for the ball. Man …
0	@nationwideclass no, it's not behaving at all …

Table 3. First 4 rows of processed data.

Sentiment	Text
0	USER URL aww that bummer you shoulda …
0	is upset that he can update his Facebook by …
0	USER dived many times for the ball man …
0	USER no it not behaving at all …

6) Removing Consecutive letters: 3 or more consecutive letters are replaced by two letters. (e.g.: "Heyyyy" to "Heyy")

7) Removing Short Words: Words with a length of less than two are eliminated.

8) Removing Stopwords: Stopwords are the English words that do not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. (e.g.: "the", "he", "have")

9) Lemmatizing: Lemmatization is the process of converting a word to its base form. (e.g., "Great" to "Good")

After preprocessing, data of Table 2 took the form of Table 3.

2.3. Results of Sentiment Analysis

Following the same steps as Chakraborty *et al.*, 5 percent of the training data from the sentiment 140 dataset was used to test the trained models. Similarly, five different models have been trained on the dataset. Namely, the models used are Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree (DT), Boosted Tree (BT), and Random Forests (RF). The 5 models have been trained with the training dataset. After this, all the models have been used to predict the sentiment value for the test data tweets. Following this step, the predicted sentiment values were compared with actual sentiment values of the test data set tweets. The results, including accuracy, precision, recall, and F1 score, are shown in Table 4.

Like the previous research report conducted by Chakraborty *et al.*, SVM performed the best when it came to overall accuracy. Logistic Regression (LR) was the next best model with a difference of 0.01 in accuracy. This was determined by subtracting the Accuracy Values of SVM and LR models.

In Table 4, precision is the equation,

$$\frac{tp}{(tp+fp)}$$

Table 4. First 4 rows of processed data.

	Accuracy	F1 Score	Precision	Recall
LR	0.82	0.82	0.82	0.82
SVM	0.83	0.83	0.83	0.83
DT	0.72	0.72	0.72	0.72
BT	0.70	0.70	0.70	0.70
RF	0.70	0.70	0.70	0.70

where *tp* is the number of true positives, and *fp* is the number of false positives. True positive would be guessing a positive sentiment when it is positive, whereas false positives would be assuming a positive sentiment when negative.

In **Table 4**, recall is the equation,

$$\frac{tp}{(tp+fn)}$$

where *fn* is the number of false negatives. A false negative is guessing negative sentiment when it is, in fact, negative. The precision is intuitively the classifier's ability not to label as positive a sample that is negative. The recall is intuitively the ability of the classifier to find all the positive examples.

In **Table 4**, the F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at one and worst score at O. The F-beta score weights recall more than precision by a factor of beta = 1.0 means recall and precision are equally important. The parameter beta = 1.0 was used in order to keep this research paper consistent with the research report conducted by Chakraborty *et al.* It is essentially a way to test the accuracy of the model.

2.4. Stock Movement Prediction

From the results of sentiment analysis, we found that SVM worked best on our test data of Sentiment 140 which is in line with the results reported by Chakraborty *et al.* For this reason, SVM was used for the sentimental analysis of the stock related tweets with the keywords, "stock market", "Stocktwits", and "AAPL". This was kept the same as the previous research study.

Tweets related to the Keyword "stock market", and "stocktwits" were trained to predict DJIA closing difference values whereas, Tweets relating to the keyword "AAPL" were introduced to predict Apple Inc. closing difference values. Tweets were ignored on the days when the Stock Market was not open. All the Keyword related tweets were preprocessed with the same process as mentioned afore. All these steps are in line with the previous work performed by Chakraborty *et al.*

2.4.1. The Score of "Stock Market", "Stocktwits", "AAPL" Related Tweets

We have used the predicted sentimental values as an output from running Sup-

port Vector Machine on the tweets corresponding to the various keywords in our work. Furthermore, we took the average sentimental value of these tweets each day and created a new dataset. We had replaced all the 0 s with −1 to have positive and negative values based on the sentimental averages. A positive average sentiment value indicates positive, while a negative average sentiment value indicates negative.

2.4.2. Stock Index Value Prediction Using Boosted Regression Tree

Similarly, to Chakraborty *et al.*, we used the Boosted Regression Tree model for the stock index value prediction. The training set is data from January to August 2016, and testing was done on stock-related data from September to December 2016. The average sentiment values of tweets containing "stock market", "stocktwits" are trained with DJIA closing price difference, while the average sentiment values of tweets containing "AAPL".

We trained our model to predict stock price differences the next day. In the training data set, the average sentiment values are of a day's tweets, and their corresponding closing price difference is between that day and the next day. So, after getting the sentiment value of tweets of the present day, we can predict how much the stock market will rise or fall the next day. In other words, for the current day's stock value prediction, we will need the previous day's tweets average marginal value.

The tweets from September to December 2016, which were used for testing went through SVM classification first, to obtain average sentiment values as our Boosted Regression Tree model is trained with average sentiment values. Our Boosted Regression Tree model used these average marginal values to predict the next day's stock difference.

2.4.3. Stock Index Value Prediction Using Multilayer Perceptron Neural Network

We improved the regression modeling by implementing a Multilayer Perceptron Neural Network model to see if neural networks better predict the stock closing difference since neural networks are shown to work better than regular models. Like the Boosted Tree model, the training set is data from January to August 2016, and testing was done on stock-related data from September to December 2016. The average sentiment values of tweets containing "stock market", "stocktwits" are trained with DJIA closing price difference, while the average sentiment values of tweets containing "AAPL".

We trained our model to predict stock price differences the next day. In the training data set, the average sentiment values are of a day's tweets, and their corresponding closing price difference is between that day and the next day. So, after getting the sentiment value of tweets of the present day, we can predict how much the stock market will rise or fall the next day. In other words, for the current day's stock value prediction, we will need the previous day's tweets average marginal value.

The tweets from September to December 2016, which were used for testing went through SVM classification first, to obtain average sentiment values as our Multilayer Perceptron Neural Network model is trained with average sentiment values. These average marginal values were then used to predict the next day's stock difference by our Multilayer Perceptron Neural Network model.

Table 5 shows the first few entries that the models were trained with.

Table 5. Shows that the average sentiment value of tweets of 01/07/2016 is 0.50, and the next day's closing price increased by 0.51 points.

Date	Average Sentiment Value of Tweets	The Actual Closing Price Difference
2016-01-04	0.50	0.51
2016-01-08	0.56	1.57
2016-01-31	0.52	0.63

3. Prediction Results of Stock Movement

We have plotted both actual stock differences and predicted stock differences in the testing period from September to December 2016. Furthermore, there are two tables below which show the Mean Average Error (MAE) and Root Mean Square Error (RMSE) between the actual stock differences and predicted stock differences by the Boosted Tree model and the MLP regression model. Table 6 shows the MAE values, while Table 7 shows RMSE values. Figure 1 shows the actual and predicted stock differences for tweets with "stock market". Figure 2 shows the actual and predicted stock differences for tweets with "stocktwits". Figure 3 shows the actual and predicted stock differences for tweets with "AAPL".

Table 6. Error Metrics MAE.

Tweets with Phrase	Mean Absolute Error of Predicting Closing Stock Price Difference with Boosted Tree	Mean Absolute Error of Predicting Closing Stock Price Difference with MLP regression
"stock market"	88.39	68.19
"stocktwits"	76.98	75.28
"AAPL"	1.37	0.98

Table 7. Error Metrics RMSE.

Tweets with Phrase	Root Mean Square Error of Predicting Closing Stock Price Difference with Boosted Tree	Root Mean Square Error of Predicting Closing Stock Price Difference with MLP regression
"stock market"	88.39	68.19
"stocktwits"	76.98	75.28
"AAPL"	1.37	0.98

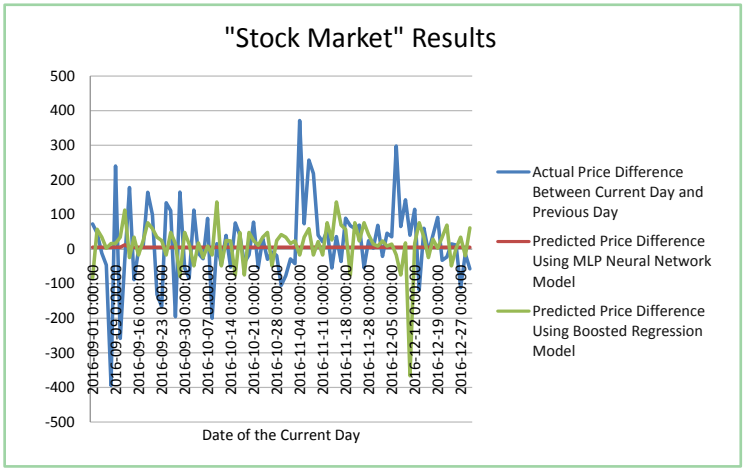

Figure 1. Prediction of closing stock price difference value using (boosted regression tree and MLP neural network) on tweets containing "stock market".

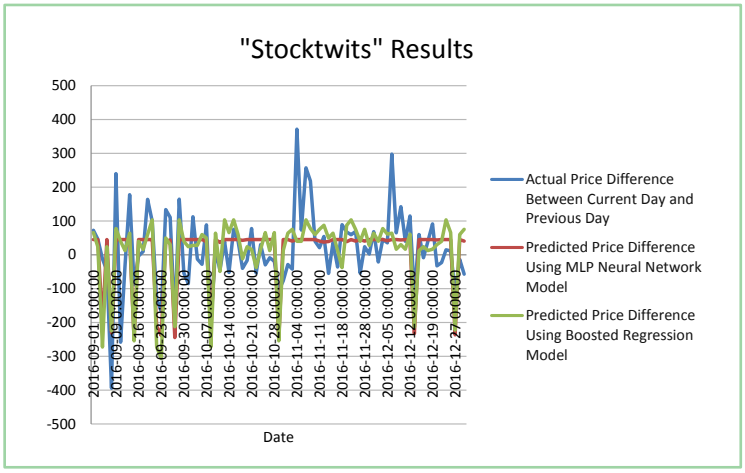

Figure 2. Prediction of closing stock price difference value using (boosted regression tree and MLP neural network) on tweets containing "stocktwits".

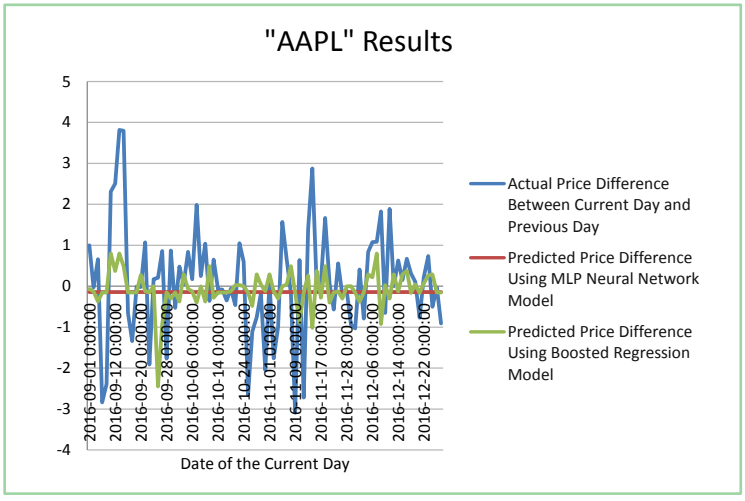

Figure 3. Company-specific ("AAPL") closing stock price difference prediction using tweets containing "AAPL" with boosted regression tree and MLP neural network.

The formula for RMSE in Table 6 is,

$$\sqrt{\frac{\sum_{i=1}^{N}(f-o)^2}{N}}$$

where Σ = summation ("add up"), $(f-o)^2$ is the difference between expected and observed values squared, and N is the sample size.

The results for each of the figures show that the MLP neural network is in fact on average better than the boosted regression tree model at predicting the Price difference of stocks. However, it is noticeable that the boosted regression tree model tended to overpredict the values whereas, the MLP neural network tended to underpredict the price difference values.

4. Conclusions and Further Work

In our work, we predict the future movement of the United States' stock market by analyzing the sentiment of Twitter posts related to the Stock market. To do this, we collected stock-related tweets and obtained their average sentiment value by using SVM. After that, we prepared the training set with those tweets and with corresponding DJIA or Apple Inc. closing stock index differences between the present-day and next day. Then we tested on similar stock related tweets on a different timeline to see how much we can predict the stock index. We used a Boosted Regression Tree model and a Multilayer Perceptron Neural Network model to do this.

We were able to derive answers for both of our hypotheses. From the results of our work, it is seen that tweets do play a role in the prediction of stock market movement. Furthermore, it is implied that Neural Networks perform better than the Boosted Regression Tree. For all three sets of data with the keywords: "stock market", "stocktwits", "AAPL", the Multilayer Perceptron Neural Network model has a lower MAE and RMSE than the Boosted Regression Tree model. From our results, it is also clear that too high and too low differences in Stock Indexes are challenging to predict with Boosted Regression Tree. However, except for those days, our models predicted very well on the given data set.

Future work regarding this study would include using the models on different stock markets across the world. Furthermore, using a data range of more than one year may provide more accurate results. Additionally, analyzing the models in different economic situations such as booms or recession may allow us to better see the productivity of the models. Besides, the use of a neural network for classifying the sentimental analysis tweets may offer better results.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Twitter Business Basics (n.d.).

https://business.twitter.com/en/basics.html

[2] Dang, L.M., Sadeghi-Niaraki, A., Huynh, H.D., Min, K. and Moon, H. (2018) Deep Learning Approach for Short-Term Stock Trends Prediction Based on Two-Stream Gated Recurrent Unit Network. *IEEE Access*, **6**, 55392–55404. https://doi.org/10.1109/ACCESS.2018.2868970

[3] Li, X., Xie, H., Chen, L., Wang, J. and Deng, X. (2014) News Impact on Stock Price Return via Sentiment Analysis. *Knowledge-Based Systems*, **69**, 14-23. https://doi.org/10.1016/j.knosys.2014.04.022

[4] Jeon, S., Hong, B. and Chang, V. (2018) Pattern Graph Tracking-Based Stock Price Prediction Using Big Data. *Future Generation Computer Systems*, **80**, 171-187. https://doi.org/10.1016/j.future.2017.02.010

[5] Agarwal, A., Xie, B., Vovsha, I.., Rambow, O. and Passonneau, R. (2011) Sentiment Analysis of Twitter Data. *Proceedings of the Workshop on Languages in Social Media LSM*'11, Stroudsburg, PA, June 2011, 30-38.

[6] Pang, B. and Lee, L. (2004) A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the* 42*nd Annual Meeting of the Association for Computational Linguistics* (*ACL*-04), Barcelona, July 2004, 271-278. https://doi.org/10.3115/1218955.1218990

[7] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. https://arxiv.org/abs/1408.5882

[8] Joachims, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Machine Learning*: *ECML*-98. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), Vol. 1398, Springer, Berlin, 137-142. https://doi.org/10.1007/BFb0026683

[9] Chakraborty, P., Pria, U.S., Rony, M.R.A.H. and Majumdar, M.A. (2017) Predicting Stock Movement Using Sentiment Analysis of Twitter Feed. 2017 6*th International Conference on Informatics*, *Electronics and Vision &* 2017 7*th International Symposium in Computational Medical and Health Technology* (*ICIEV-ISCMHT*), Himeji, 1-3 September 2017, 1-6. https://doi.org/10.1109/ICIEV.2017.8338584

[10] Khatri, S.K. and Srivastava, A. (2016) Using Sentimental Analysis in Prediction of Stock Market Investment. IEEE 5*th International Conference ICRITO*, Noida, 7-9 September 2016, 566-569. https://doi.org/10.1109/ICRITO.2016.7785019

[11] Chen, R. and Lazer, M. (2011) Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement. Cs 229, pp. 15.

[12] Khan, W., Ghazanfar, M.A., Azam, M.A., Karami, A., Alyoubi, K. and Alfakeeh, A. (2020) Stock Market Prediction Using Machine Learning Classifiers and Social Media, News. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-020-01839-w

[13] Kaz Anova, M.M. (2017) Sentiment140 Dataset with 1.6 Million Tweets. https://www.kaggle.com/kazanova/sentiment140

[14] Yahoo Finance—Stock Market Live, Quotes, Business & Finance News (n.d.). https://finance.yahoo.com/