STOCK PRICE PREDICTION USING MICROBLOGGING DATA

SP Lakshmi Narayanan[1], R Kamalendran[2], Mohit kumar[3], C Murale[*]

[1,2,3]Student, Department of IT, Coimbatore Institute of Technology, Tamilnadu, India
[*]Assistant Professor, Department of IT, Coimbatore Institute of Technology, Tamilnadu, India

---***---

Abstract - *Stock market prediction is that the act of trying to work out the longer term value of a corporation stock held at some securities market. Successful prediction of a stock's future price could yield a major profit. Volatile market makes prediction difficult. "Stock Price Prediction Using Microblogging data" a technique for predicting stock prices is developed using news articles and historical stock prices. The changes available prices of a corporation, the rises and falls, are correlated with the general public opinions being expressed in tweets this company. Understanding author's opinion from a chunk of text is the objective of sentiment analysis. Positive news and tweets in social media a couple of company would definitely encourage people to speculate within the stocks of that company and as a result the stock price of that company would increase. A prediction model for locating and analyzing correlation between contents of tweets and stock prices so making predictions for future prices may be developed by using machine learning. stock market could be a subject that's highly littered with economic, social, and political factors. There are several factors e.g. external factors or internal factors which may affect and move the exchange. Stock prices rise and fall every second thanks to variations in supply and demand. Techniques using machine learning will give more accurate, precise and easy thanks to solve such issues associated with stock and market prices so use of regression method provides better prediction.*

***Key Words*: stock market analysis; sentiment analysis; tweets; microblogging; prediction**

1. INTRODUCTION

The stock exchange is essentially an aggregation of varied buyers and sellers of stock. A stock (also known as shares more commonly) in general represents ownership claims on business by a particular individual or a group of people. The attempt [3] to determine the future value of the stock market is known as a stock market prediction. The prediction is expected to be robust, accurate and efficient. The system must work according to the real-life scenarios and should be well suited to real-world settings. The system is also expected to take into account all the variables that might affect the stock's value and performance. There are various methods and ways of implementing the prediction system like Fundamental Analysis, Technical Analysis, Machine Learning, Market Mimicry, and Time series aspect structuring. With the advancement of the digital era, the prediction has moved up into the technological realm. The most prominent and [3] promising technique involves the implementation of machine learning. Machine learning involves AI which empowers the system to find out and improve from past experiences without being programmed time and again. Traditional methods of prediction in machine learning use algorithms like Backward Propagation, also referred to as Backpropagation errors. Lately, many researchers are using more of ensemble learning techniques. It would use low price and time [3] lags to predict future highs while another network would use lagged highs to predict future highs. These predictions were used to form stock prices. [1].

Although, technical data is vital for stock prediction contemporary traders need more advanced strategies to outperform market. According to behavioural economics it be could useful to add information about emotions, moods and psychological states of people [4]. Though uninteresting individually, Twitter messages, or tweets, can provide an accurate prediction of public sentiment on when taken in aggregation. In this paper, we primarily examine the effectiveness of varied machine learning techniques on providing a positive or negative sentiment on a tweet corpus.

We have used implemented techniques in two different ways: 1) To predict tweet sentiment score based on classification co-efficient and 2) The correlation between tweet sentiment and stock market prices.

2. METHODOLOGY

2.1. TWITTER SENTIMENT ANALYSIS

Microblogging data have characteristics which will indicate potential informative value to the forecasting of stock exchange behavior. Services like Twitter and StockTwits have large communities of investors sharing information about stock exchange . These users frequently interact during the day and react readily to events. Messages are usually very objective thanks to the character limit. Microblogging users usually apply cashtags available market conversations to ask the involved stocks. Cashtags are composed by a "$" character and therefore the respective ticker (e.g., $AAPL) and its presence means the message is said thereupon stock. the use of cashtags permits a simple and fewer noisy selection of messages associated with specific stocks. The extraction of attention and sentiment from microblogging is quicker and cheaper than from traditional sources (e.g., surveys) because data is promptly available at very low cost.

The sentiment and a spotlight indicators created during this study were extracted from Twitter, which may be a large microblog platform with quite 300 million active users. Using Twitter REST API (https://dev.twitter.com/docs/api), we collected all messages holding cashtags of all stocks traded in US markets.

There are 6 objective for sentiment analysis
1.Collect social data from twitter
2.Process and clean tweet corpus
3.Generate feature set and data set from processed tweet
4.Analyze and predict tweet sentiment using classification techniques
5.Compute tweet mood and generate data set for stock prediction
6.Predict stock exchange trend from twitter sentiment statistics..

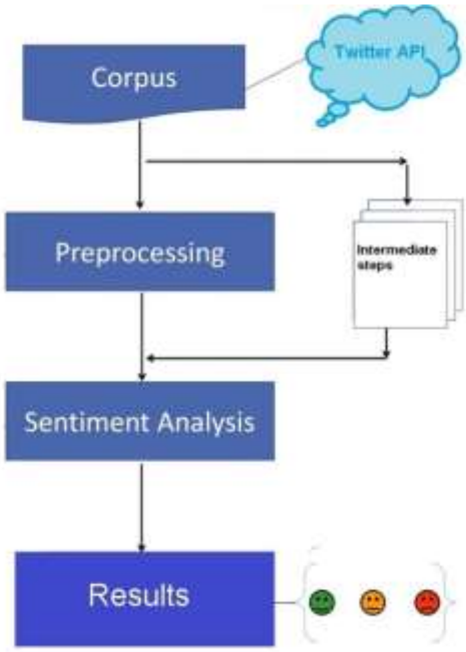


Fig1. Sentiment analysis model

- For tweet collection, Twitter provides robust API to gather tweets: The streaming API or the search API. To overcome the limitation of streaming API, we have used search API. The search API is REST API which allow users to request specific query of recent tweets.
- The search API allow more fine tuning queries filtering based on time, region, language etc.
- The request of JSON object contains the tweet and their metadata. It includes variety of information including username, time, location, retweets. We have focused on time and tweet text for further analysis purpose.
- An API requires the user have an API key authentication. After authenticating using key, we were able to access through python library called Tweepy.
- The text of each tweet contains too much extraneous words that do not consider to its sentiment. Tweets include URLs, tags to others and many other symbols that do not have any meaning. To accurately obtain tweet's sentiment we need to filter noise from its original state.
    - An example of tweet characters and formats can be seen in below table

- TABLE I. TWITTER EXAMPLE

Company Name	Tweet status/Text
$AAPL	@JeffMacke Stunning that they never did this. Or $AAPL, or $TWX- these guys all should have been trying to buy $NFLX.
$GOOG	Oracle, #Google fail to settle Android lawsuit before retrial. Read more: https://t.co/oYX31p9CDZ $GOOG
$YHOO	Investors hope Yahoos sale puts check in the mail $YHOO https://t.co/S6DviHMrfm https://t.co/4mUngBgHzt https://t.co/czUcEPLrjO
$MSFT	StockTwits: EARNINGS Mon - $IBM $PEP $NFLX Tue - $JNJ $GS $YHOO Wed - $KO $AXP Thu - $UA $GOOG $SBUX $MSFT $GM Frâ€¦ https://t.co/eaFaIcF9Wr
$GS	Oil Prices to Fall Asr Producers Fail to Reach Deal https://t.co/IvYFfYIeR7 via @WSJ Commodities head Jeff Currie $GS called it last week

- First step is to split the text by space, forming a list of individual words per text which is called as list of words. We will use each word in tweet as feature to train our classifier. Next step is to remove stop words from list of total words.
- Stop words contains articles, punctuation and some extra words which do not have any sentiment value. It contains stop word dictionary which check each word in list of tokenized words against dictionary. It word is stop word then it filters it out.

- Tweets mostly contain symbols like @, #, $, URLs, extra spaces and punctuations for different purposes. All the symbols except $ are removed because they add no value in sentiment analysis. The words start with $ is ticker of company name so we can filter out them as they may contain useful information for sentiment analysis.
- After processing, we excluded the unnecessary columns and prepared a CSV file that contains three attributes i-e time, tweet text and sentiment.
- As large amount of tweets data have been collected for training purpose, now we can build and train the classifier. We have used SVM because it is fairly robust to overfitting, it can handle large feature spaces and it is memory efficient. Further, in previous works, SVMs have been shown to be very effective for text categorization. First the whole data was split into training and test set. To create the training and test set with 80% training set probability. Then the training set was used to train the model and both training and test accuracies were calculated to evaluate the model.
- As we know there could be more than one tweet on each day about one company, so the data that we got from sentiment prediction in previous is aggregated day-wise. It means on a day if there are more positive tweets than negative we say stock sentiment is positive on that day and a person can buy the share. After training our classifier, we decide to move on an application to look at correlation between tweet sentiment and stock market prices on each day scale. To do so, we have collected stock data as well as tweet data for same timeline. In addition, we focus on specific company stocks and gathered day by day data for each.

2.2. PREDICTION METHODS

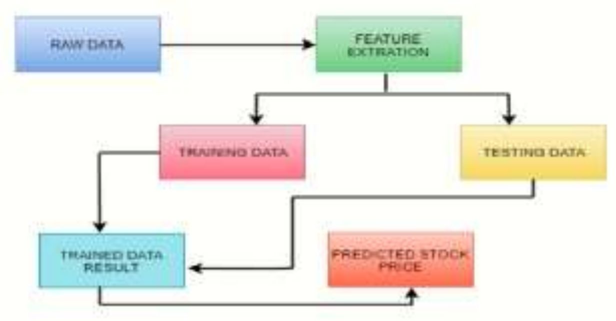


Fig2. Prediction model

2.2.1. Linear regression

If the goal is prediction, or forecasting, or error reduction, linear regression are often wont to fit a predictive model to an observed data set of y and X values. After developing such a model, if a further value of X is then given without its accompanying value of y, the fitted model are often wont to make a prediction of the value of y. Regression predicts a numerical value [10]. Regression performs operations on a dataset where the target values are defined already. And the result can be extended by adding new information [10]. The relations which regression establishes between predictor and target values can make a pattern. This pattern are often used on other datasets which their target values aren't known. Therefore the info needed for regression are 2 part, first section for outlining model and therefore the other for testing model. In this section we elect rectilinear regression for our analysis. First, we divide the info into two parts of coaching and testing. Then we use the training section for starting analysis and defining the model.

2.2.2. K-Nearest Neighbor

The focus of regression problems is to predict the output based on variables given by a set of independent variables. k-Nearest Neighbors Regression (kNR) is completed by simply assigning the property value of the thing to be the average of its *k* nearest neighbors. It will be useful to weight the contributions of the neighbors so that the nearer neighbors contribute more to the typical than the more distant ones[6]. kNR makes predictions supported on the result of the *k* neighbors closest thereto point. Therefore,

$$D(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (1)$$

to make predictions with kNR, a metric to measure the distance between the query point and cases from the example samples is needed [6]. One of the foremost popular choices to measure this distance is that Euclidean distance. Euclidean distance is suitable for calculating numerical predictions. Eq. (1) is the formula for Euclidean distance [6]. where D is distance between p and q, n is amount of data, and i is the index i After selecting the value of k, we can make predictions based on the kNR. For regression, the predictions are the average of the k nearest neighbors outcome. Eq. (2) is the formula to make the predictions [6].

$$y = \frac{1}{k}\sum_{i=1}^{k} y_i \qquad (2)$$

where y is the predicted value, k is the number of neighbours, yi is the stock price value at index i.

2.2.3. Long Short Term Memory

Long Short-Term memory is one among the foremost successful RNNs architectures. LSTM introduces the memory cell, a unit of computation that replaces traditional artificial neurons within the hidden layer of the network. With these memory cells, networks are ready to effectively associate memories and input remote in time, hence suit to understand the structure of knowledge dynamically over time with high prediction capacity.

In this stage, the information is fed to the neural network and trained for prediction assigning random biases and weights. Our LSTM model is consists of a sequential input layer followed by 2 LSTM layers and dense layer with ReLU activation then finally a dense output layer with linear activation function[7]. The output value generated by the output layer of the RNN is compared with the target value.

The error or the difference between the target and therefore the output value is minimized by using back propagation algorithm which adjusts the weights and the biases of the network.

2.2.4. Random Forest

Random Forest is an ensemble classification algorithm that uses a collection of decision tree in combination. Random Forest was first introduced by Leo Breiman following on the ideas of Amit et al and Ho . The method requires the random selection of features (or attributes) to split at each of the decision tree node. The random factor makes the individual trees uncorrelated. This makes the Random Forest[5] robust to noise and resistant to over training. Each of the trees, at the end of the tree traversal, will cast a vote for the classification of the input class; the sum of the total vote that constitutes the majority will be the classification. A single random tree classifier will only have a slightly better than random classification but combining them as an ensemble can produce very much improved accuracy. A feature of Random Forest is that it does not overfit but will reach a limiting value of generalization error.

3. ANALYSIS

For analyzing the efficiency of the system we are used the Root Mean Square Error (RMSE). The error or the difference between the target and therefore the output value is minimized by using RMSE value. RMSE is that the root of the mean/average of the square of all of the error. The utilization of RMSE is highly common and it makes an superb general purpose error metric for numerical predictions. Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum\left(\hat{Y}_i - Y_i\right)^2}$$

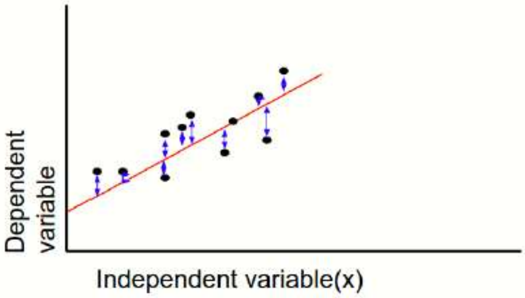


Fig 3: RMSE value calculation

4. EXPERIMENTAL RESULTS

We have implemented application in number of phases. We would like to describe our results and discuss about results in following manner.

In first phase of task, we have collected stock tweets using twitter API. It responses tweets in JSON format. It is mandatory to keep configuration file as an authentication to download tweet data live.

The indispensable configuration parameters are:

```
{
  "consumer_key":"",
  "consumer_secret":"",
  "access_token":"",
  "access_token_secret":""
}
```

After fetching twitter corpus, we have store it in text file. In parallel we process our tweets for further analysis purpose.

To process tweet, we have follow some basic steps:

1. Tokenized tweets
2. Remove extra keyword from tweets
3. Remove stop words from tweet
4. Replace URLs and meaningful notation with meaning full keywords

To prepare training set we have used AFINN library. It computes tweet keywords positive and negative scores and result tweet sentiment with positive, negative or neutral score. We have map tweet score with positive, negative or neutral keyword. If total tweet score is greater than 0 then it considers as positive keyword. If total tweet score is less than 0 then it considers as negative. If total tweet score is 0 then it is considered as neutral.

For example, the processed tweet is “Oracle, #Google fail to settle Android lawsuit before retrial”. The total sentiment score of this tweet -4. Hence it is considered as negative tweet. The generated tweet corpus csv file contains tweet date, tweet text, tweet sentiment keyword and total tweet score. We have prepared dataset using generated tweet corpus csv file.

Total available tweets for five different companies (15thApril):

Table III. Company tweet counts

COMPANY	POSITIVE	NEGATIVE	NEUTRAL	TOTAL
AAPL	25	14	61	100
GOOG	28	8	64	90
YHOO	35	4	61	100
MSFT	20	8	72	100
GS	17	15	68	98

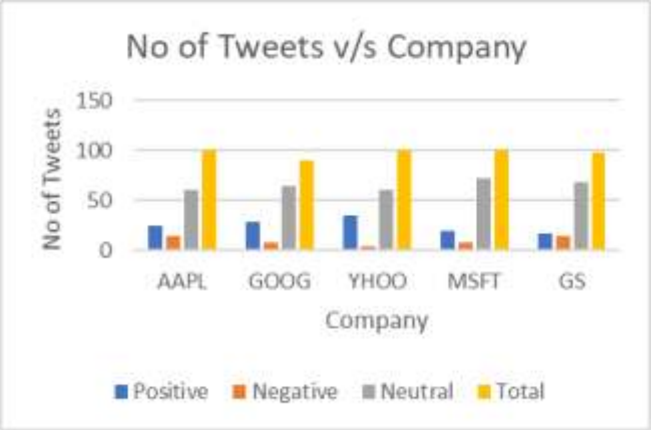


FIGURE 4. NUMBER OF TWEETS VS COMPANY

Approximately, we can fetch around 90-10 tweet of each stock company for a single day. The above graph represents relation between company stock name and total number of tweets (positive, negative, neutral and total). Using Python's csv library, we have read tweet corpus data which will converted into model input dataset. Tweet corpus file provide data in four different columns. We read tweet text from csv file and create feature set by choosing important keyword from whole dataset. The feature set will compare with each tweet and found present feature in tweet. If feature is presented, then it will consider as feature 1. If feature is absent, then it will be considered as feature 0. Using stated, approach we have built feature matrix. It considers feature matrix in form of 0 and 1 and target matrix in form of positive, negative and neutral. For example, tweet is "Oracle, #Google fail to settle Android lawsuit before retrial". The important keywords are fail and lawsuit so they are considered as 1 and other keywords are considered as 0.

We have also captured tweet statistics and stock prices for last eight days. For more accuracy we do not process weekend dataset. For Apple company stock prediction, we have collected tweet corpus using $APPL keyword and one other hand we have collected yahoo finance data for APPL stock prices.

The Stock price data is got using yahoo finance API used to download the dataset for the given company name and the date range. The stock price data is stored in a Comma Separated Values (CSV) file.

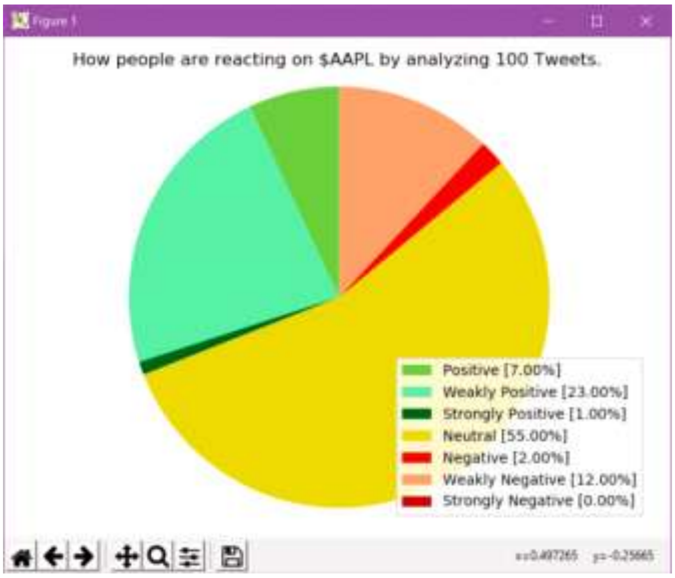
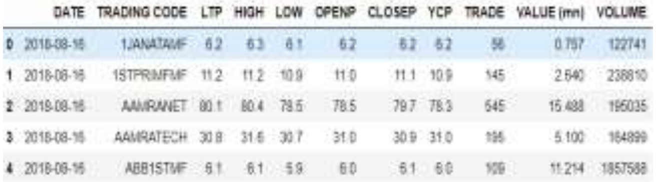


Fig:5 Sentiment analysis of AAPL

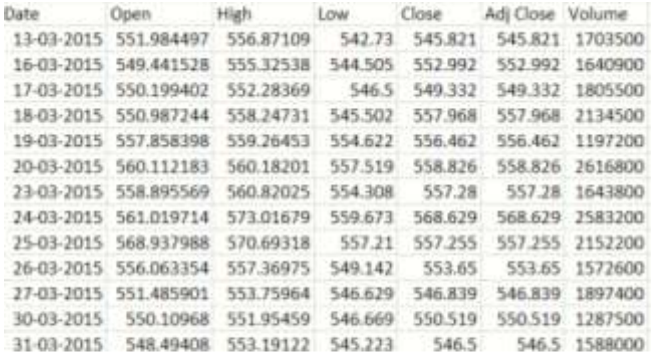
We Retrieved the last five years stock price data of GOOG (Alpabet Inc).



	DATE	TRADING CODE	LTP	HIGH	LOW	OPENP	CLOSEP	YCP	TRADE	VALUE (mn)	VOLUME
0	2018-08-16	1JANATAMF	6.2	6.3	6.1	6.2	6.2	6.2	56	0.757	122741
1	2018-08-16	1STPRIMFMF	11.2	11.2	10.9	11.0	11.1	10.9	145	2.640	238810
2	2018-08-16	AAMRANET	80.1	80.4	78.5	78.5	79.7	78.3	545	15.488	195035
3	2018-08-16	AAMRATECH	30.8	31.6	30.7	31.0	30.9	31.0	196	5.100	164899
4	2018-08-16	ABB1STMF	6.1	6.1	5.9	6.0	6.1	6.0	109	11.214	1857588

Fig:6 Price Head

Usually There are Seven columns or seven attributes that describe the rise and fall in stock prices. Some of these attributes are (1) HIGH, which describes the highest value the stock had in previous year. (2) LOW, is quite the contrary to HIGH and resembles the lowest value the stock had in previous year. (3) OPENP, is the value of the stock at the very beginning of the trading day and (4) CLOSEP stands for the price at which the stock is valued before the trading day closes. There are other roles that are not needed for our finding such as YCP, LTP, TRADE, VOLUME and VALUE.



Date	Open	High	Low	Close	Adj Close	Volume
13-03-2015	551.984497	556.87109	542.73	545.821	545.821	1703500
16-03-2015	549.441528	555.32538	544.505	552.992	552.992	1640900
17-03-2015	550.199402	552.28369	546.5	549.332	549.332	1805500
18-03-2015	550.987244	558.24731	545.502	557.968	557.968	2134500
19-03-2015	557.858398	559.26453	554.622	556.462	556.462	1197200
20-03-2015	560.112183	560.18201	557.519	558.826	558.826	2616800
23-03-2015	558.895569	560.82025	554.308	557.28	557.28	1643800
24-03-2015	561.019714	573.01679	559.673	568.629	568.629	2583200
25-03-2015	568.937988	570.69318	557.21	557.255	557.255	2152200
26-03-2015	556.063354	557.36975	549.142	553.65	553.65	1572600
27-03-2015	551.485901	553.75964	546.629	546.839	546.839	1897400
30-03-2015	550.10968	551.95459	546.669	550.519	550.519	1287500
31-03-2015	548.49408	553.19122	545.223	546.5	546.5	1588000

Fig.7 Raw stock price data

This is the Pictorial representation of the data present in our csv file. This particular file contains records for 5 years prices. Some of the records do not have relevant information that can help us train the machine, so the logical step is to process the raw data. Thus we obtain a more refined dataset which can now be used to train the machine. Start training the data and split the data into 80% for Training and 20% for Testing the more the data we provide for training the mode accuracy we get. After training the data with all the algorithms start testing and plot the graph using "matplotlib.pyplot" library. The plot is based on CLOSEP and DATE after plotting calculate Root Mean Square Value (RMSE). The model with less RMSE value Provides the best accuracy. After the sentiment analysis if the sentiment score is positive and the price is increasing the user is recommended to buy the shares. If the sentiment analysis score is negative and the price is decreasing then the user is recommended to sell the shares soon.

5. CONCLUSION

On random walks and numerical prediction but with the introduction of behavioural finance, the people's belief and mood were also considered while predicted about stock movement. Making it more efficient we used the idea of sentiment analysis of Stock Tweets through machine learning models. We implemented the idea by collecting sentiment data and stock price market data and built multiple models for prediction and in the last we measured the prediction accuracy. Results showed that we have achieved better accuracy in LSTM Model compared to other models. It can be improved more if we increase the size of data set. The models did not perform well in cases where stock prices are low or highly volatile. There are still different ways to build stock prediction models, which we leave as future work. Some of these include building a domain-specific model by grouping companies according to their sector, considering adverse effects on the stock price of a company due to news about other related companies, and considering more general industry and global news that could indicate general market stability. There is need to improve the test accuracy of text classification algorithm. More training data can be used to enhance the prediction accuracy. This can be extended to support decision making for more investment options like commodity market and real state.

ACKNOWLEDGEMENT

We are immensely grateful to Mr.Murale C, Assistant Professor, Dept. of Information Technology, Coimbatore Institution of Technology, Coimbatore for his constant guidance and unending support.

REFERENCES

[1] Bing Li,Keith C.C Chan,Carfol Ou,Sun Ruifeng, "Discovery public sentiment in social media for predicting stock movement of publicly listed companies," Elsevier,Information Systems 69C pp. 81–92,2017,

[2] S.Kokila, Dr.A.Senthilrajan , Implementation of Extended Deep Neural Networks for Stock Market Prediction,Proceedings on International Journal for research in Applied Science and Engineering Technology(IJRASET), *Volume 8 Issue II, Feb 2020*,.

[3] Bo Zhao,Yonji He,chufeng Yuan, "Stock market prediction exploiting microblog sentiment analysis," Proceedings of the International joint conference on Neural network ,pp 4482-4488, 2016

[4] Sushree das,Ranjan Kumar Behera,Mukesh kumar,Santanu Kumar Rath, "Real Time Sentiment Analysis of Twitter Streaming data for Stock Prediction," Proceedings of the International conference on Computational Intelligence and Data Science(ICCIDS),pp 956-964, 2018

[5] Loke K.S, "Impact of financial ratios and technical analysis on stock price using random forests" Procedings of the International conference on Computer and Drone Applications(IConDA),pp 38-42,2017

[6] Julius Tanuwijaya,Seng Hansun, "LQ45 Stock Index Prediction using K-Nearest Neighbours Regression,"vol. 8, pp. Issue-3, September 2019.

[7] Murutaza Roondiwala, Harshal Patel, Shraddha Varma, "Predicting stock prices using LSTM,"Procedings of the International Journal of Science and Research(IJSR),vol. 6, Issue-4 ,pp 1754-1756, April 2017.

[8] Xue Zhang,Hauke Fuehres,Peter A. Gloor, "Predicting stock market indicators through twitter,",Proceedings of Collaborative Innovation Networks Conference, pp. 55-62, 2011.

[9] K.Hiba Sadia,Aditya Sharma, Adharrsh Paul, "Stock market prediction using machine learning algorithms,",Proceedings of International Jornal of Enginnering and Advanced Technology(IJEAT),Volume-8 Issue-4,April 2019.

[10] R.Seethalakshmi,”Analysis of stock market predictor variables using Linear Regression”,Proceedings of International Journal of pure and applied mathematics,Volume 119 No.15,pp 369-378,2018