Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 441 – 449

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

Prediction Models for Indian Stock Market

Aparna Nayak, M. M. Manohara Pai* and Radhika M. Pai

Manipal Institute of Technology, Manipal University, Manipal 576 104, India

Abstract

Stock market price data is generated in huge volume and it changes every second. Stock market is a complex and challenging system where people will either gain money or lose their entire life savings. In this work, an attempt is made for prediction of stock market trend. Two models are built one for daily prediction and the other one is for monthly prediction. Supervised machine learning algorithms are used to build the models. As part of the daily prediction model, historical prices are combined with sentiments. Up to 70% of accuracy is observed using supervised machine learning algorithms on daily prediction model. Monthly prediction model tries to evaluate whether there is any similarity between any two months trend. Evaluation proves that trend of one month is least correlated with the trend of another month.

Keywords: Boosted Decision Tree; Logistic Regression; Sentiment Analysis; Stock market; Support Vector Machine.

1. Introduction

Stock price prediction is very important as it is used by most of the business people as well as common people. People will either gain money or lose their entire life savings in stock market activity. It is a chaos system. Building accurate model is difficult as variation in price depends on multiple factors such as news, social media data, fundamentals, production of the company, government bonds, historical price and country's economics[1]. Prediction model which considers only one factor might not be accurate. Hence incorporating multiple factors news, social media data and historical price might increase the accuracy of the model.

There are two common methods to predict the stock market prices[2]. One among that is chartist or technical theories and the second one is fundamental or intrinsic value analysis. Proposed method is built on the principle of technical theories. Basic assumption of this theory is history tends to repeat itself. Prediction model can be applied on the historical data to get future trend. As researchers have discussed in S. J. Grossmara and R. J. Shiller[3] and L. Andrew and M. A. Craig[4], as and when new information comes in the market stock market value varies. Technical analysis and semi strong form of efficient market hypothesis are followed, to build prediction model in the proposed work. The goal of this research work is to build a model which predicts stock trend movement (trend will be up or down) using historical data and social media data. Two models are built as part of research work. Both models use supervised machine learning algorithm. First model is daily prediction model, considers both sentiment and historical data.

*Corresponding author. Tel.: +91-9945202361.
E-mail address: mmm.pai@manipal.edu

doi:10.1016/j.procs.2016.06.096

This model predicts the future trend for the next day. Sentiment of the company has been computed by using twitter data and news of the company. Outcome of sentiment analysis is considered along with open price, close price of stock with extracted statistical parameters to build model. Second model is monthly prediction model, considers only historical data and predicts the trend for next one month. Proposed work 3 investigates whether the outcome of model is inline with the actual trend movement.

The rest of this paper is organized is as follows. Section 2 introduces some previous research work on sentiment analysis for stock market prediction and stock trend movement using historical price. Section 3 describes proposed method. Section 4 shows the dataset used and evaluates the results of the experiments. Finally, Section 5 concludes the contribution of this research work.

2. Related Work

Many research groups are exploring stock market trend prediction using social media analytics. Architecture for building the model has been referred from P. Paakkonen, D. Pakkala[5]. Many use case architecture have been discussed in the same paper. Multiple methods are there to detect the polarity of each tweet/news. Initially moods of a user on the specific company was considered to analyze the stock price as shown in X. Zhhang *et al.*[6] and J. Bollen *et al.*[7]. Now polarity of each item in news/tweet has been found to get the sentiment. To find the polarity of each news/tweet item one can use either dictionary based approach or semi supervised algorithm. In case of dictionary based algorithm, polarity to each word is assigned by comparing each word of news with dictionary word. In case of semi supervised algorithm as discussed by K. Mizumoto *et al.*[8], initial level of dictionary is built manually then new words are categorized as either positive or negative based on occurrence of new words along with words in the built dictionary. Dictionary based approach has been used in the proposed method, as semi supervised learning might not cover all possible combination of words.

X. Zhang *et al.*[6] and J. Bollen *et al.*[7] have analyzed that mood of individual affects stock market price. In[6] they have also mentioned that twitter sentiment might effect stock market trend only for few company. W. Antweilwer and M. Z. Frank[9] have discussed that information which is available about any company is not noise. One can get useful information such as prediction of future value from it. R. Ahuja *et al.*[10] have analyzed twitters on stock market by collecting 3 months BSE data. N. Lin *et al.*[11] have shown that news effects future market trend of stock market. They have considered two market places America and China. M. Hagenau *et al.*[12] have considered German Adhoc messages as input and for feature selection, Chi square method has been used. SVM algorithm has been used for which 65% of accuracy is obtained.

J. Gong and S. Son[13] have implemented stock prediction model using logistic regression considering feature index variables. They have mentioned that daily stock trading prediction with logistic regression out performs other methods such as RBF – ANN prediction model.

3. Proposed Method

The model predicts the price movement on t_n by considering all the available historical data i.e. from $t_{n-1}, t_{n-2}, \dots t_1$, where t_n stands for transaction data of prediction. All the available data is trained by supervised machine learning algorithm. Sentiment from social media data and news are extracted. Extracted sentiments later will be integrated with historic price to build prediction model. Conflicting opinions has been reported by researchers about effect of sentiment on stock market. Few research[14] reported sentiment extracted from social media has no effect on stock price movement whereas in[7], they have reported the sentiment has either strong or weak effect on stock price movement.

Two different models have been built to predict stock market trend. First model predicts the stock market trend for the next day (Daily prediction model) by considering all available data on daily basis as input. Second model predicts the stock market trend for the next month(Monthly prediction model) by considering available data on monthly basis.

First contribution of the proposed work is that few features has been deduced from the historical data available by using statistics. One of the statistical parameter considered is relationship between trend of a day and volume of stock traded on the same day[15]. Volume traded feature in historical data will reflect both bought and sold stocks on a daily

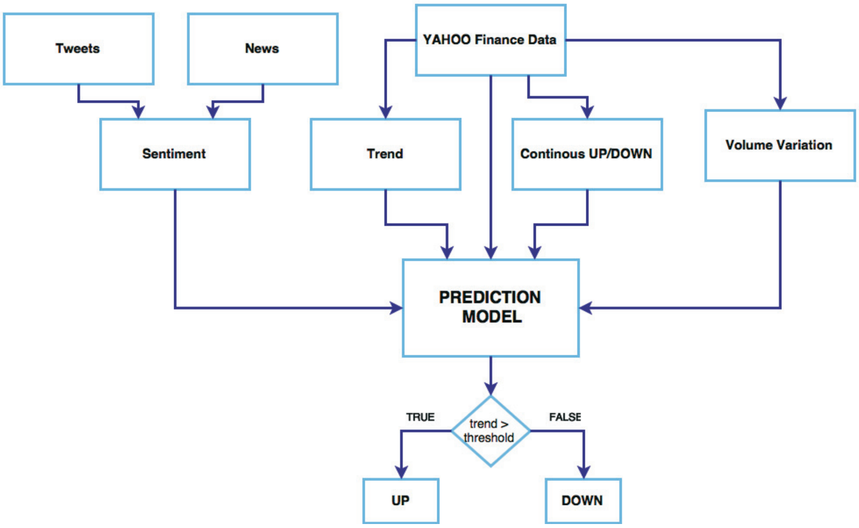

Fig. 1. Prediction Model for Daily Prediction Model.

basis. When the trend is up volume traded might indicate the sold shares, similarly when the trend is down volume traded might reflect shares bought by traders. This feature has been combined with trend of that day to get whether the volume of stocks is sold or bought by the trader. Big number of volume traded has positive impact if and only if shares are purchased by the trader. Assumption for the shares purchased is stock transactions are more and trend is down. If volume traded is more and trend is up means, shares are sold to gain money. One more statistical parameter is computed by considering past n days pattern of up/down. These features are generated for training and testing dataset. Now the prediction model is built on training dataset. Another contribution of this paper is Monthly prediction model. In this the entire month trend is computed by considering historical data. Input to the model is given month wise. Month $m's$ and year $y's$ prediction is based on year $m-1, m-2, \ldots$ of year y. Here the assumption is trend of month m in the year y will follow trend of some different month in the same year.

3.1 Daily prediction model

Daily prediction model is built by considering historical price dataset and sentiment dataset. Prediction model for the same is as shown in Fig. 1.

Figure 1 shows input and output of the prediction model. Patterns like continuous up/down, volume analysis has been derived from Yahoo finance data. Sentiment from news dataset and tweets are also given as input to the model. Prediction model is trained by using supervised machine learning algorithm. Model is tested on test data, which tells whether threshold is up or down.

3.1.1 Sentiment analysis

The following steps has been followed for the sentiment analysis[16].

- **Data collection:** Data is gathered using following methods.
  - News information is collected from 2 different websites using crawler.
  - Tweets are collected using twitter API using python language.
- **Data Processing:** Collected data is processed using following methods.
  - **Lemmatization:** Lemmatization is applied to each row to get the all words to common form which will be helpful while assigning polarity to each word. Lemmatization is done with the help of natural language tool kit(NLTK) package which is available in python.

- **Data Analysis:** Collected data has to be analyzed to get the sentiment on each day.
  - **Assigning polarity:** The data is moved on hadoop distributed file system. Using hive query, polarity has been assigned to each word by comparing existing dictionary. Polarity of each item is calculated by summing up the polarity of each word which appears in the news/tweet item.

Nomenclature

P	Set of positive words
N	Set of positive words
U	Set of all users
R	Set of all user tweets/news
W	Set of all words
r	Tweet/news of any user $u \in U$
u	User $u \in U$
w	Any word in the review/news $\in W$

```
Procedure followed to assign polarity for tweets:
    $\forall_u \ u \in U$
    if $\exists_r \ r_w \in R \ \& \ w \in r \ \& \ w \in P$
        $pol_w = 2$
    else if $\exists_r \ r_w \in R \ \& \ w \in r \ \& \ w \in N$
        $pol_k = -1$
    else $pol_k = 0$
```

3.1.2 Historical price analysis

Yahoo finance data has been analyzed, feature transformation has been applied to get new features like continuous up/down, relationship between volume traded and trend are found.

- **Pattern 1 – Continuous up/down:** Closed price variations having continuous up/down for 5 days is considered as a pattern. In step 1 of Algorithm 1, rend on each day is calculated. It is calculated be subtractingtoday's close price from yesterday's close price. If obtained value is non negative number then trend is positive otherwise negative. In step 2 of Algorithm 2 continuous up/down has been discussed. Continuous up/down is calculated by considering trend of last three days.
- **Pattern 2 – Volume variation:** In the data provided by YAHOO, stock volume traded on daily basis is available. Volume traded on each day is compared with trend on the same day to get volume variation pattern as shown in Algorithm 3.

```
Step 1 - Trend calculation
Data: Close price vector(cV)
Result: Trend on each day(t)
cV = difference(cV);
append 0 to cV;
j=0;
for i in cV do
    if $i > 0$ then
        t[j] = 1;
    else
        t[j]=0;
    end
    increment j;
end
```

Algorithm 1. Algorithm to Calculate Trend on Each Day

```
Step 2 - Identification of Continuous five days up/down
Data: Date vector along with trend vector(d)
Result: Vector containing relationship between trend and volume traded(t)
i = 0;
j=0;
for i,i+1,i+2,i+3 in d do
    for j,j+1,j+2,j+3 in t do
        if j == j+1 && j==j+2 && j==j+3 then
            cD=1;
        else
            cD = 0
        end
    end
end
```

Algorithm 2. Algorithm to Check Continuous Days Up/Down

```
Data: Vector containing volume traded on each day (vV) and Vector containing trend on each day (t)
Result: Vector containing relationship between trend and volume of stock traded(vD)
vV = difference(volume) ;
append 0 to vV ;
for i in vV do
    for j in t do
        if $i == i - 1$ then
            vD=0;
        else if $i > i - 1$ && j==1 then
            if $i > (2 * (i - 1))$ then
                vD = 1;
            else
                vD = 0.5;
            end
        else if $i < i - 1$ && j==0 then
            if $2 * i < (i - 1)$ then
                vD = -1;
            else
                vD = -0.5;
            end
        else
            vD = -1
        end
    end
end
```

Algorithm 3. Algorithm to Find the Relationship Between Trend and Volume of Stock Traded

The sentiment found in section 3.1.1, is combined with patterns found in section 3.1.2 to build the prediction model. Algorithm to combine the sentiment and the historical data is given as in Algorithm 4. Date of sentiment is compared with date in the stock price variation dataset. If there is a match data is combined.

The data generated is classified as training and testing dataset. 60% of total data is considered as training dataset and rest data is considered as testing dataset. The prediction model is built by considering training dataset. Testing dataset is used to test the accuracy of the model.

3.2 Monthly prediction model

Monthly prediction model is built by considering the whole month data as one data point. So the correlation between each month is calculated. Correlation between each month trend is calculated using hamming distance measure. Fig. 2 shows how the correlation between different months has been compared.

```
Data: Data frame consisting of historical data along with volume variation pattern and continuous up/down
      (reqData)
Result: Combined data frame (combainedData)
for i in reqData[date] && j in senti[date] do
    if $i == j$ then
        combainedData[i] = reqData[i] ;
    else if $i < j$ then
        increment i;
    else
        increment j;
    end
end
```

Algorithm 4. Algorithm to Combine Sentiment with Historical Data

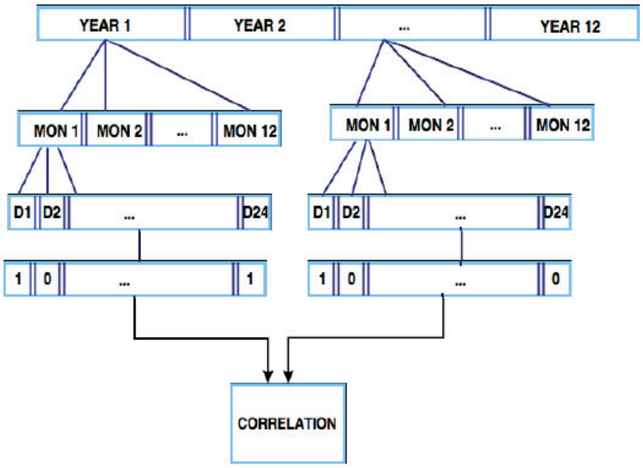

Fig. 2. Correlation Model.

Correlation between each month's of historical data from year 2002 to 2014 as well as six months of 2015 is calculated. The model is built by considering the data of 2015 as label. It is assumed that there are 24 trading days in a month.

Let b is monthly trend of any month which is a vector of 24 bits. P_{ik} and P_{jk} be the pattern at any month $i \& j$ and at position k where $i \neq j$

$$\text{correlation} = \frac{\Sigma_k b_k}{|P_{ik}|} \quad \text{where} \quad b_k = \begin{cases} 1 & \text{if } P_{ik} = P_{jk} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Correlation which is discussed in equation 1 has a positive value over the historical dataset. Data from 2003 to 2014 is considered as features and 2015 data is considered as label. Total 6 models are built using logistic regression model. Each model is trained with month wise data. July month's data is tested on each month's model.

Entire month trend is given as input to train the model. Algorithm to convert the entire data frame into month wise data is as shown in Algorithm 5. Month wise data i.e. a pattern of 24 values is trained, to predict trend for one complete month.

The idea is to predict the whole 24 bit pattern trend for any month by considering previous months 24 bit pattern of trend.

```
Data: Data frame consisting of historical data along with volume variation pattern and continuous up/down (hD)
Result: Month wise data for prediction
pM = 0;
tM = Vector to hold 24 days trend ;
p = Data frame which consists of multiple months trend ;
hD = Data frame consisting of historical data ;
for i in 1:number of rows(hD) do
    if pM == 0 then
        pM = month(i)
    if month(i) == pM then
        append t(i) to tM;
    else
        assign tM to p;
    end
end
```

Algorithm 5. Algorithm to Divide the Entire Dataset into Month wise

4. Experimentation and Results

In the proposed method two datasets are used. The first one is historical price dataset, and second one is sentiment dataset of that company. Sentiment has been calculated from the obtained news and tweets information of the company. Companies are selected in such a way that it is from different sector. One of the company is from oil sector, another is from bank sector and last one is from mining sector. For daily prediction model, last one year's tweets and historical data is considered. It has almost 260 rows of data. For monthly prediction model, historical data from 2003 has been collected.

4.1 Dataset

4.1.1 Historical prices

Historical prices are obtained from Yahoo Finance. Each transaction date consists of open price, close price, low price, high price, adjusted close price and volume traded on that day. Adjusted close price and close price depicts the close price of stock on a particular day. Adjusted close price will be adjusted for dividends and splits. Adjusted close price is considered as stock price as in other researches[17, 18].

4.1.2 Sentiment dataset

Sentiment dataset has been created by considering news dataset and tweets. Both tweets and news has been collected for one year and sentiment analysis algorithm is applied on the same. Sentiment of the tweets and sentiment of the news are integrated on a daily basis.

The classification models are built for stock market data analysis. Performance of the model is evaluated using accuracy metric. Accuracy can be defined as proportion of true results in the test dataset.

4.2 Results

4.2.1 Daily prediction model

Label has been assigned to each transaction date by considering close price on a particular day with close price of the previous day. Total observation per company is 261 transaction dates. In these, 182 transaction dates data are used for training the model. The rest 79 transactions are used for testing the model. The graph in Fig. 3 depicts original trend versus predicted trend for all three sectors. The Table 1 shows the accuracy achieved in the daily prediction

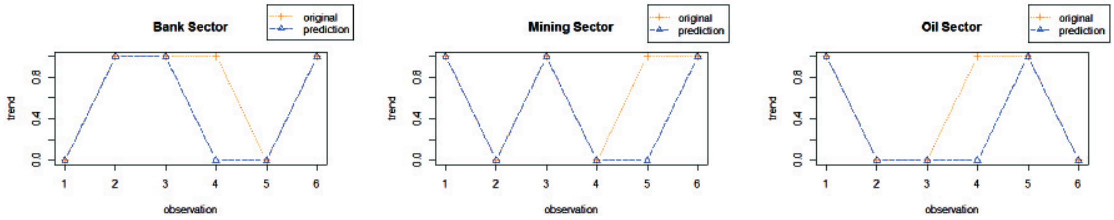

Fig. 3. (a) Oil Sector; (b) Mining Sector; (c) Bank Sector.

Table 1. Accuracy of Monthly Prediction Model.

Model used	Bank	Mining	Oil
Boosted Decision Tree	0.548	0.76	0.769
Logistic Regression	0.654	0.61	0.442
Support Vector Machine	0.51	0.59	0.442

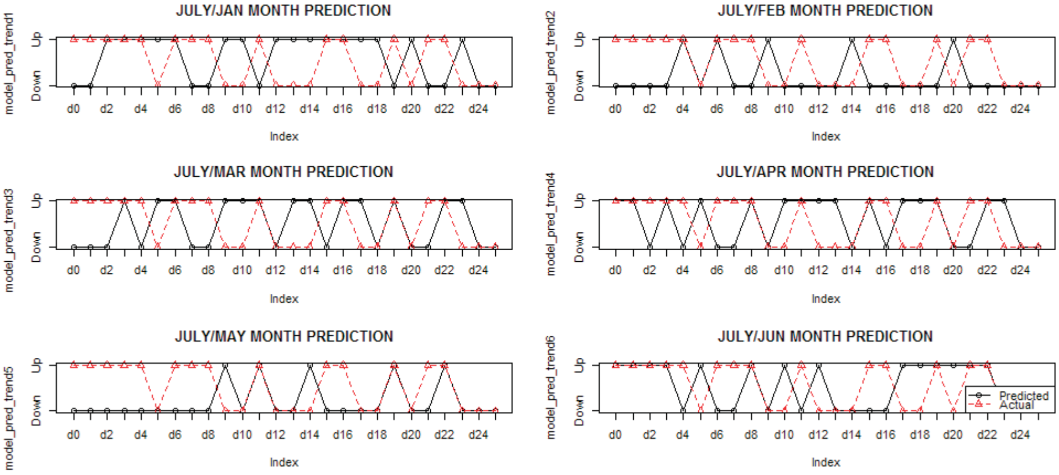

Fig. 4. July Month's Prediction by Considering January – June Month's Model.

model using three different supervised machine learning algorithms. As we can see accuracy of the model is high when boosted decision tree is used and it is low for support vector machine.

4.2.2 Monthly prediction model

Data from 2003 to 2015 has been collected. Model is trained with supervised machine learning algorithm month wise. Data from 2003 to 2014 is considered as features, 2015 data is considered as label to train the model.

As shown in the graph (Fig. 4) we can see that july month's trend is in line with the previous month's trend. When the experiment is repeated for other data it is observed the one month's data need not to be in line with its previous month.

5. Conclusions and Future Work

In the past few years, it has been observed that most of the people are investing in the stock market to make money easily. At the same time investor has high chance of losing all money invested. So an efficient predictive model is required for the user to understand future market trend. There are many predictive models which tell about the market trend whether it is up or down, but they fail to give accurate results. An attempt has been made to build efficient predictive model of stock market where the trend for the next day is predicted. By considering various patterns like continuous up/down, volume traded per day and also including sentiment of the company a model has been built and tested with different stock market data available open source. On the considered dataset, Decision Boosted Tree is performing better than Support Vector Machine and Logistic Regression.

The dataset which was been considered for sentiment analysis may be sparse which means we may not have news/tweet for a particular company for many days. In such cases Principle component analysis with multiple factors can be applied. The impact of intra day price movement for the next day stock price can be considered to improve the accuracy. Monthly prediction model can be made more accurate by considering sentiments. Value of correlation is high between two months, one can identify the news/tweet items to get the common issue during those month.

References

[1] S. A. R. Nai-Fu Chen and Richard Roll, Economic Forces and the Stock Market, *The Journal of Business*, vol. 59, no. 3, pp. 383–403, (1986). [Online]. Available: http://www.jstor.org/stable/2352710.

[2] E. F. Fama, Random Walks in Stock Market Prices, *Financial Analysts Journal*, vol. 51, no. 1, pp. 75–80, (1995). [Online]. Available: http://www.jstor.org/stable/4479810.

[3] S. J. Grossman and R. J. Shiller, The Determinants of the Variability of Stock Market Prices, *National Bureau of Economic Research*, Working Paper 564, October (1980) [Online]. Available: http://www.nber.org/papers/w0564.

[4] A. W. Lo and A. C. MacKinlay, Stock Market Prices do not Follow Random Walks: Evidence from a Simple Specification Test, *Review of Financial Studies*, vol. 1, no. 1, pp. 41–66, (1988).

[5] P. Pääkkönen and D. Pakkala, Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems, *Big Data Research*, vol. 2, no. 4, pp. 166–186, (2015). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214579615000027.

[6] P. A. G. Xue Zhang and Hauke Fuehres, Predicting Stock Market Indicators through Twitter I Hope it is not as Bad as I Fear, *Procedia – Social and behavioral Sciences*, vol. 26, pp. 55–62, (2011).

[7] J. Bollen, H. Mao and X. Zeng, Twitter Mood Predicts the Stock Market, *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, (2011). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187775031100007X.

[8] K. Mizumoto, H. Yanagimoto and M. Yoshioka, Sentiment Analysis of Stock Market News with Semi-Supervised Learning, In *2012 IEEE/ACIS 11th International Conference on Computer and Information Science (ICIS)*, pp. 325–328, May (2012).

[9] M. Z. F. Werner Antweiler, Is all that Talk Just Noise? the Information Content of Internet Stock Message Boards, *The Journal of Finance*, vol. 59, no. 3, pp. 1259–1294, (2004). [Online]. Available: http://www.jstor.org/stable/3694736.

[10] R. Ahuja, H. Rastogi, A. Choudhuri and B. Garg, Stock Market Forecast Using Sentiment Analysis, In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1008–1010, March (2015).

[11] N. Lin, J. Yuan, W. Xu, L. Wei and X. Wang, How web News Media Impact Futures Market Price Linkage?, In *2013 Sixth International Conference on Business Intelligence and Financial Engineering (BIFE)*, pp. 562–566, November (2013).

[12] M. Hagenau, M. Liebmann, M. Hedwig and D. Neumann, Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Specific Features, In *2012 45th Hawaii International Conference on System Science (HICSS)*, pp. 1040–1049, January (2012).

[13] J. Gong and S. Sun, A New Approach of Stock Price Prediction Based on Logistic Regression Model, In *2009. NISS '09. International Conference on New Trends in Information and Service Science*, pp. 1366–1371, June (2009).

[14] R. F. W. Robert Tumarkin, News or Noise? Internet Postings and Stock Prices, *Financial Analysts Journal*, vol. 57, no. 3, pp. 41–51, (2001). [Online]. Available: http://www.jstor.org/stable/4480315.

[15] G. W. Schwert, Why does Stock Market Volatility Change Over Time? *The Journal of Finance*, vol. 44, no. 5, pp. 1115–1153, (1989). [Online]. Available: http://dx.doi.org/10.1111/j.1540-6261.1989.tb02647.x.

[16] G. V. Attigeri, M. P. M. M, R. M. Pai, and A. Nayak, Stock Market Prediction: A Big Data Approach, In *TENCON 2015 - 2015 IEEE Region 10 Conference*, pp. 1–5, November (2015).

[17] P. S. Michael Rechenthin and W. Nick Street, Stock Chatter: Using Stock Sentiment to Predict Price Direction, *Algorithmic Finance*, (2013).

[18] T. H. Nguyen, K. Shirai and J. Velcin, Sentiment Analysis on Social Media for Stock Movement Prediction, *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611, (2015). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417415005126.