Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Stock Market Prediction with Historical Time Series Data and Sentimental Analysis of Social Media Data

Kesavan M, Karthiraman J, Ebenezer Rajadurai T and Adhithyan S
Department of Computer Technology, MIT Campus,
Anna University,
Chennai, India.
ebenezerrajadurai5@gmail.com

***Abstract*—In the Indian stock market, stock costs are viewed as exceptionally fluctuating due to various factors such as political decision results, bits of gossip, budgetary news, public safety events and so on. This fluctuation behavior makes it a difficult and challenging task to predict stock prices. The proposed work aims to develop a new methodology by combining sentiment analysis along with normal stock market prediction from time-series data using deep learning techniques. It extracts sentiments from news events, social media platforms particularly from twitter and incorporates the polarity of the sentiments to enhance the prediction accuracy. From the results and analysis, it can be observed that the proposed approach improved the forecasting accuracy after introducing the sentiment polarity scores. Hence, this model helps the investors to make a wiser and better decision about their investment.**

Keywords—Stock Market, Stock price, Time Series data, Sentiment Analysis, LSTM.

I. INTRODUCTION

A stock market is a public market to sell or buy the shares of a company or an organization. The ownership of the company is divided into smaller units called shares or equity or stocks that provide the rights to share in the profit of the company. When an organization needs capital for its development or expansion it issues shares called Initial Public Offer (IPO) under the primary share market. The person who invests in the stock of a company is termed as the stockholder of that company. The shares bought in the primary market can be sold in the secondary market. All these activities are facilitated by stock exchanges. The two primary stock exchanges in India are the National Stock Exchange (NSE) and the Bombay Stock Exchange (BSE) which are regulated by the Securities and Exchange Board of India (SEBI). The index of NSE is NIFTY – 50 and that of BSE is SENSEX. The stock movements and fluctuations of the NSE stock price variation of a company over a particular time are depicted in fig 1.

Stock market prediction is a tough and challenging activity since the stock prices are directly influenced by numerous factors. The factors include political and government decisions, budgetary news, public safety events, etc. Public safety events may be man-made events like terrorist activities or natural disasters or it may be due to some epidemic. Recently, the outbreak of global pandemic novel Coronavirus (COVID 19) has greatly affected the stock market all over the world. In India, the trading activity had been stopped for about 45 minutes in an unprecedented manner. So, it is necessary to predict these fluctuations as early as possible so that the investors can make wiser decisions.

Fig. 1. Fluctuations of stock price

At present web, information is increasing rapidly due to the increase in a number of social media sites such as Facebook, YouTube, Twitter, Instagram, forums, blogs, and reviews. The number of users using these platforms is also increasing. This web content contains so many useful information such as public safety events, government policies, decisions, political events, investors' opinions, etc. These events and sentiments have a direct effect on the fluctuations in the stock market. It is needed to extract much useful information from a large amount of unstructured data. This can be done with the help of Natural Language Processing (NLP) techniques such as sentiment analysis. Hence, the objective of this paper is to improve the prediction accuracy of the stock market movement by combining sentiments and emotion information of various events extracted from the social media along with the normal time-series data-based prediction.

The rest of the paper is organized as follows. Section II elaborates various related models for stock price prediction. The detailed architecture and working mechanism of the proposed methodology for stock market prediction is given in section III. Section IV includes the experimental setup and result in analysis. Finally, conclusion and future scope of the work is presented in Section V.

II. RELATED STUDY

Hiransha.M. et. al. compare 4 deep neural networks namely Recurrent Neural Networks (RNN), Multi-Layer Perceptron (MLP), Convolution Neural Networks (CNN) and Long Short

Term Memory units (LSTM) and concludes CNN performs well among other models [1]. The prediction can be further improved by a hybrid network that combines other models. Two variants of CNN were introduced by Hoseinzade E. et. al. in which one variant can be applied to data collected from single source while the other variant can be applied to data collected from various markets [2]. But CNN is not suitable for time series prediction since its prediction does not depend on previous outcome. Zhang K. et. al., proposed Generative Adversarial Networks (GAN) based technique to predict close price of the stock where the generator is built using LSTM and discriminator is built using MLP [3]. Further, this model can be improved by using suitable optimizer and including other factors that affect financial data.

The work presented by Shi. L, et. al., performed training a deep neural network for text based prediction, visualizing patterns and evaluate with real-life scenario [4] and it can be enhanced by the interpretation of other social media messages along with financial news events. A prediction approach by combining autoencoder, deep learning model and restricted Boltzmann machine was introduced by Chen. L, et. al., and found that the performance is better than existing machine learning approaches such as extreme learning machine, radial basis function neural network and back propagation neural network [5]. By incorporating other factors like politics, economy, culture, environment, etc., the model can be enhanced. Zhang. X. et. al., presented a novel data extraction method to extract data from various sources like web news events and social media events and developed a multiple instance machine learning model for classification [6]. The resources other than twitter messages can be included for sentiment analysis to enhance the predicted value accuracy.

A deep learning based multi feed neural network was proposed by Long W. Lu. Z, et. al. to extract features from multivariate historical financial time-series data and compared the results with RNN and CNN [7]. To further enhance the performance, other factors that affect stability and profitability can be explored. A hybrid framework based on machine learning techniques for stock index price prediction is introduced by Chen.Y, et. al., in which weights were assigned to features during training phase using Support Vector Machine (SVM) and during testing phase features with weights were used by K-nearest neighbor (KNN) algorithm [8]. Other correlation weighting methods and other hybrid machine learning models are to be considered for further improvement.

Four machine learning models, SVM, ANN, naïve-Bayes and random forest were used by Patel.J, et.al. with two different approaches for input data [9]. Only the closing price of a stock is predicted by them. Relative strength, stochastic oscillator, and moving average convergence were calculated. The prediction can be further improved by ensemble of various machine learning methods. An attempt was made by Idrees. S. M, et. al., to design linear statistical model to forecast future stock price using the Auto-Regressive Integrated Moving Average (ARIMA) model which is a combination of Auto-Regressive (AR) and Moving Average (MA) models [10]. But this model is only suitable for forecasting univariate time-series data. To minimize the error rate the enhancement of the ARIMA model with some hybrid model can be done.

An algorithm that combines motif-based sequence reconstruction along with Convolutional neural network to predict the stock price was proposed in [11]. Sentiments from news feed can be added to improve accuracy. Numerical-based Attention (NBA) methodology was proposed by Nousi P, and Tsantekidis A which uses dual information namely news and numerical information for stock prediction [12]. The NBA was used with machine learning and deep learning models. A novel Two-stream Gated Recurrent Unit (TGRU) was proposed in [13] which uses both sentiment data as well as financial data. But the complexity of TGRU is doubled that of GRU and takes more training time and computational resources. SVM based prediction method was proposed in [14] which uses sentiment data extracted from news events to incorporate investor psychology. To handle a large volume of data deep learning-based models can be used.

III. PROPOSED METHODOLOGY FOR STOCK MARKET PREDICTION

The proposed method combines two varieties of inputs for enhanced stock market prediction. One is data extracted from twitter and news events. The raw data extracted from social media platforms contain noisy data. Hence, pre-processing is necessary to remove irrelevant data. From the preprocessed data, the psychology of the investors and the impact of news events are identified by natural language processing techniques which gives polarity scores for the news and social media data.

The second input is financial time series data extracted by web crawling. The suitable deep learning method for analyzing time series data is RNN.

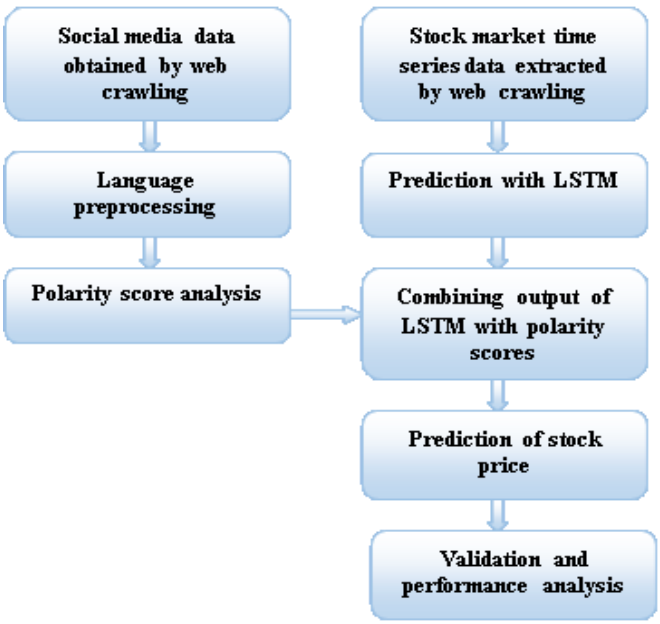

Fig. 2. The detailed architecture of the predictor model

Since RNN is suffered from vanishing and exploding gradient problem, LSTM is used to analyze the financial time series data. Now, the challenge is to incorporate the sentiments extracted from news and social media events to the financial time series data prediction. A simple algorithm has been

proposed to associate the sentiments and time-series data prediction. The detailed flow of activities is given in Fig. 2.

Finally,combined model is evaluated and the performance of the system is analyzed. The entire process is divided into four major activities which are explained below.

A. Data Extraction and Preprocessing

Two different varieties of data are extracted using web crawling and manual crawling methods. The first variety of data is historical stock time-series data. It has attributes like opening price, closing price, minimum value of stock, maximum value of stock and shares traded. These data can be intraday data, day-wise data, weekly data, or monthly data. This historical stock data needed to be normalized before directly applying the values to the model since the parameters used were of different scales. Two normalization techniques are commonly used:

1) Min-max scalar

This converts the prices in the range 0 to 1 which is given in the equation (1).

$$Normalized\ Value = \frac{(Current\ Value - Min\_value)}{(Max\_value - Min\_value)} \quad (1)$$

where Max_value, Min_value are the maximum and minimum value of the data to be normalized. This technique does not retain information about prices increasing or decreasing as compared to the previous day.

2) Percentage change from previous value

This converts the prices in the range -1 to 1 which is given in equation (2). After normalization, negative normalized value indicates a decrease in stock price and positive normalized value indicates an increase in stock price. Min-max scalar method gave better results than change in percentage method which is given in Algorithm 1.

Algorithm 1: Min-Max Scalar calculation

```
1. Extract dataset of a company containing stock prices of a
   particular company:
2. Normalized value = 0
3. Max_Value = 0
4. Min_Value = 0
5. For every data D of the dataset:
6.    Normalized_value = Current Value – Min_Value
7.    Max_Value = Max_Value - Min_Value
8.    NormalizedValue = Normalized_value / Max_Value
9.    If Normalized_value < 0:
10.         stock prices decrease
11.   Else if Normalized_value>0:
12.         stock price increases
13. End
```

$$Normalized\ value = \frac{(Current\ Value - Previous\ Value)}{Previous\ Value} \quad (2)$$

The second variety of input data is data extracted from news headlines and twitter. Data pre-processing of financial headlines and twitter data is essential to remove noises such as stop words, redundant words, and punctuations as they are not important for financial and sentiment analysis. Punctuations and characters were removed, contractions were replaced by their same form by using a dictionary expanding English language contractions in python [15]. NLTK in python provides a set of stop words in English language in corpus module which can be used to remove stop words from the news and social media data. News headlines and tweets were first tokenized and stop words can be removed by comparing the list of stop words in corpus module. To convert the inflectional forms of words to the base form stemming was performed. For example, “banking” is converted to “bank”.

B. Sentiment Analysis of Social Media Data

Natural Language Preprocessing (NLP) is a branch of artificial intelligence that deals with interaction between human and computer using natural language. It is the process of reading, understanding, decoding human-understandable natural language such as text, speech, signs, etc. The field of NLP includes making computers to perform valuable errands with the characteristic dialects people use. Breaking and tokenize the sentence, predicting parts of speech, finding noun phrases and identifying semantic meaning are some of the challenges in NLP. The steps of natural language processing are given in Fig. 3.

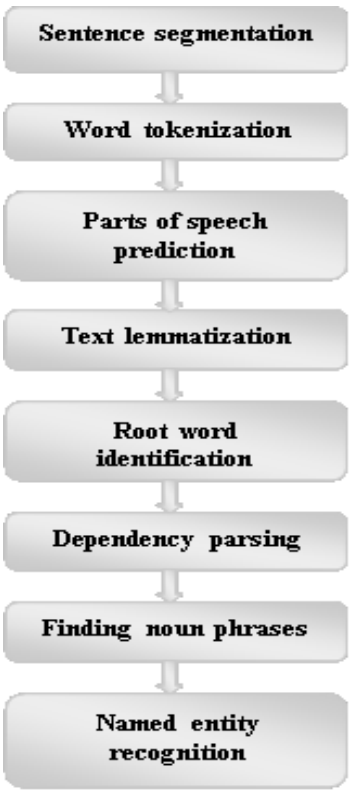

Fig. 3. Steps of Natural Language Processing

Sentiment analysis is the process of extracting and interpreting emotions from a piece of text using text mining

techniques. It is mainly used to identify the emotions and opinions of the customers about a service or product. In our case, the customers are investors of the stock. Detection of sarcasm, negations, word ambiguity and multipolarity are some of the challenges in sentiment analysis.

Sentiment Analysis technique analyses the social media message and classifies the sentiment of the message into three categories namely positive, negative or neutral. The sentiment analysis first tokenizes the sentence and then remove unwanted words. Then for each word the polarity of the word is calculated based on the lexicon. Lexicon is the set of words along with their emotion. Based on the emotion the polarity score is calculated. The polarity score calculation is given in Algorithm 2.

C. *Prediction with LSTM on Time Series Data Input:*

To analyze sequential data RNN is used since it uses output at time 't' to compute output at time 't+1'. Due to this feature RNN is used in various applications like language translation, image captioning, speech recognition, language modeling, etc. But, the problem with RNN is vanishing and exploding gradient problem. LSTM and GRU overcome this problem and hence suitable for analyzing historical stock time-series data. LSTM also solves the problem of handling long-term dependencies which cannot be handled by RNN.

Algorithm 2: Polarity score calculation

```
1. Extract dataset of a company containing news events
   about company:
2. Make google glove's dictionary as reference:
3. Positive_score = 0
4. Negative_score = 0
5. Neutral_score = 0
6. For every news N of dataset:
7. Do
8.     Tokenisation and stemmatization [ news_text]
9.     For every word in news_text:
10.        If word in positive_text:
11.            Positive_score += 1
12.        Else if word in negative_text:
13.            Negative_score += 1
14.        Else:
15.            Neutral_score += 1
16. End
```

LSTM can be productive in securities exchange expectations as it equipped for adapting long haul conditions. Stock costs of earlier days and earlier years are to be kept in memory to effectively anticipate a result. The key to LSTM is the cell state. LSTM maintains both long-term and short-term memories which is represented by cell state. It is also capable of adding or removing information from the cell state that is regulated by a structure called gates. Forget gate takes input from the previous state and decides whether to keep the information or to throw away the information using sigmoid function. Input gate is used to update the cell state. It uses sigmoid function and tanh function to update the information. The cell state is calculated by pointwise addition of output from forget gate and the output form input gate. The output gate decides which information should be passed to the next hidden state which controlled by sigmoid and tanh function. Inclusively Adaptive Momentum (ADAM) optimizer is used because of its performance on the fluctuating data. In addition to that, ADAM optimizer is direct to execute, computationally productive, invariant to corner to corner rescale of the inclinations, suitable for non-stationary targets appropriate for issues with scanty angles. Hence little memory is needed for this Adaptive Momentum optimizer.

In the proposed work, the LSTM model takes earlier day's stock costs with opening, most extreme, least and last exchanging costs as info parameters. Information was a rate change. LSTM network comprised of 4 information neurons with every neuron speaking to a solitary parameter from the information dataset. A think back worth was set for preparing the model. This esteem speaks to the quantity of lines which ought to be thought back for anticipating the following value i.e. on the off chance that lookback esteem is set to 3, at that point model ought to consider the stock prices of past 3 days to foresee the following day's cost. The model was prepared on different lookback values. The yield of the model was rate change in the closing cost of the following day's stock.

D. *Combining Sentiments with Time Series Prediction*

In this module, the output from LSTM and sentiments from social media data is combined to predict whether there is a chance of increase or decrease of stock price is given in Algorithm 3. First, from the data set of LSTM model (real time series data), set a threshold point of both upper and lower threshold point, through which a saturation point is obtained by considering the lower threshold point. Then, analyze the sentiments from twitter account, financial news events and other social media's public quoted data and identify their overall polarity score in terms of positivity rate, neutral rate and negativity rate. Finally, on preposition logic given in algorithm 3, concluded that it was going to be increase or decrease in the future and with the prediction of the close price of the very next working day.

Algorithm 3: Combining polarity scores with LSTM output

```
1. Start
2. pos= sum (positive sentiments from twitter)
3. neg = sum (negative sentiments from twitter)
4. pos1= sum (positive sentiments from news events)
5. neg1= sum (negative sentiments from news events)
6. predic_value= sum of all predictions
7. predic_average= predic_value / total number of
       predictions
8. If pos > neg and pos1 > neg1 and pred_avg >
   Lower_thresh and Pred_avg > Higher_thresh:
9.         stock price will increase
10. Else if pos  > neg  and pos1 > neg1 and pred_avg >
   Lower_thresh:
11.        stock price will stagnant
12. Else:
13.        Stock price will decrease
```

IV. EXPERIMENTAL SETUP AND PERFORMANCE ANALYSIS

A. Dataset

The dataset for the dedicated company is obtained from MoneyControl [16]. MoneyControl provides a dataset for training models for real time stock market. This data is being utilized for training LSTM model. This dataset provides complete data of the company which can be used for the users to predict the hike/decline of stock price in future. Each time the program runs automatic web crawling takes place from moneycontrol. The dates are also flexible and can be changed according to user's wishes. Similarly, the company can also be chosen by the user. In this case, Infosys datais taken and predicted its stock price. In addition to that using the twitter data, news event data and some facebook data (posts) for enhancing the prediction accuracy. These data are collected through both manual and web crawling.

B. Tools used

1) NumPy

It is a package for performing scientific computation with Python [17]. It provides tools for performing data analytics and also provides provisions for graphical analytics.

2) Pandas

Pandas is a powerful tool built using python for data analysis and data manipulation [18]. The main functionalities of Pandas include data alignment, data integration, data reshaping and handling various file formats.

3) Natural Language Toolkit (NLTK)

Nltk is a fast and flexible toolkit for natural language processing which is commonly used for manipulation of human understandable text and speech by software [19].

4) Keras

Keras is a high-level API for deep neural network modelling written using python and runs on the top of Theano or Tensorflow [20]. It is simple, easy to use, modular and composable.

5) Matplotlib

Matplotlib is an extension of numpy for producing high quality figures, charts and graphs [21]. The graphs and images can be downloaded in various hardcopy formats.

6) Scikit-Learn

Scikit-learn is a famous module in python for machine learning [22]. The commonly used features are classification, clustering, dimensionality reduction, ensemble methods and cross validation.

C. Results and Analysis

Fig. 4 and Fig. 5 shows the positive and negative polarity scores respectively obtained from news and social media events for Infosys.

The forecasting of stock prices from time-series data is shown in Fig. 6 and Fig. 7. Fig 6 shows the training and testing of the outcome of closing price prediction of the stock. From the figures, it can be observed that the predicted stock values are almost close to the actual values. Thereby we can conclude that LSTM can be able to predict the stock market efficiently.

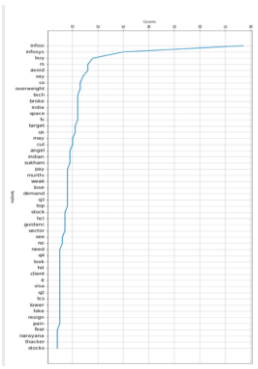

Fig. 4. Positivity analysis of social media data

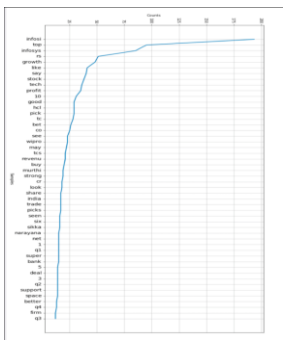

Fig. 5. Negativity analysis of social media data

The result of sentiment polarity scores is then incorporated along with the times series data prediction. Fig. 8 shows the predicted result whether the stock price will increase or decrease and what will be the expected predicted price based on the sentiment features.

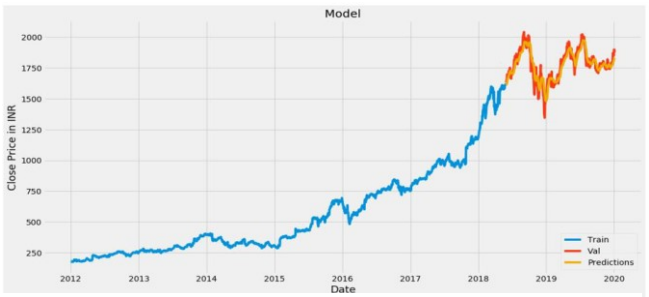

Fig. 6. Training and testing of time series data using LSTM

Date	Close	Predictions
2018-05-31	1629.619995	1610.071411
2018-06-01	1641.540039	1615.097290
2018-06-04	1665.270020	1621.031982
2018-06-05	1696.349976	1629.084839
2018-06-06	1695.750000	1640.332153
...	...	...
2019-12-27	1869.800049	1788.256348
2019-12-30	1846.890015	1800.353271
2019-12-31	1847.839966	1810.291382
2020-01-02	1898.010010	1818.331665
2020-01-03	1874.969971	1829.339966

[402 rows x 2 columns]
[[1838.3601]]

Date	
2019-12-17	1790.660034
2019-12-18	1784.030029
2019-12-19	1792.280029
2019-12-20	1786.500000
2019-12-23	1793.000000
2019-12-24	1789.209961
2019-12-26	1868.770020
2019-12-27	1869.800049
2019-12-30	1846.890015
2019-12-31	1847.839966
2020-01-02	1898.010010
2020-01-03	1874.969971

Name: Close, dtype: float64

Fig. 7. Actual and predicted values of close stock

The performance of the proposed stock price prediction method can be analyzed based on the parameters such as error percentage, accuracy, and precision. Percentage error can be calculated from actual value and the predicted value as given in equation (3).

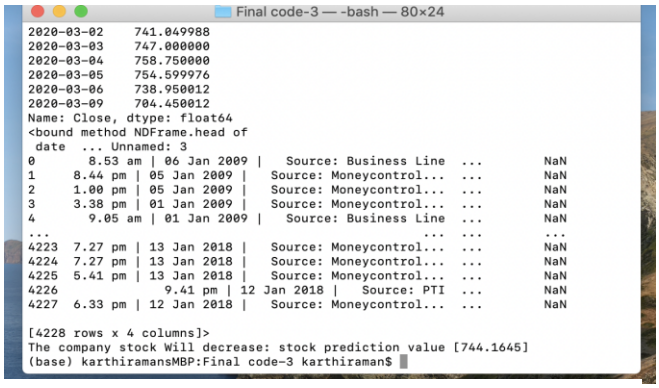

Fig. 8. Predicted result after incorporating sentiment polarity scores

$$\%\,error = \frac{Actual\ close\ price - predicted\ close\ price}{Actual\ close\ price} X\ 100 \qquad (3)$$

Accuracy is calculated by subtracting percentage error from 100 which is shown in equation (4).

$$Accuracy = 100 - Percentage\ Error \qquad (4)$$

Precision means a fraction of relevant instances among retrieved instances that can be calculated based on the difference between the average actual close price and average actual predicted close price which is given in equation (5).

$$Precision = \frac{Average\ actual\ value - Average\ predicted\ value}{Average\ predicted\ value} \qquad (5)$$

Based on the above metrics, the percentage error of the proposed method is 3.05 percentage and hence the accuracy of the proposed system is about 96.95 percentage which is greater than the existing works. Thus, our sentimental analysis and times series data prove to be very much useful and thereby increasing its accuracy.

V. CONCLUSION AND FUTURE WORK

The proposed work presents a precise technique for stock market prediction in the perspective of Indian economy. It uses sentiment analysis to extract polarity scores from the news and social media content and incorporates the extracted sentiments along with historical stock time-series data to forecast the stock price. Since, the events and psychology of the investors have a direct influence on the stock market, the proposed method yields accurate results. The percentage error of the proposed method is about 3.05 which is lesser than other existing methods. Hence, it helps the investors to make wiser decisions during various events that affect the stock market.

In future, sentiments from other social media platforms can be included to improve the performance of the proposed system. Additionally, by consolidating more stocks from various spaces and discovering the relationship among them to anticipate the pattern for a better prediction.

REFERENCES

[1] Hiransha. M, Gopalakrishnan. E. A, Menon. V. K and Soman. K. P, "NSE stock market prediction using deep-learning models", Procedia computer science, Vol. 132, pp. 1351-1362, 2018.
[2] Hoseinzade. E and Haratizadeh. S, "CNNpred: CNN-based stock market prediction using a diverse set of variables", Expert Systems with Applications, Vol. 129, pp. 273-285, 2019.
[3] Zhang K., Zhong, G., Dong, J., Wang, S. and Wang, Y., "Stock Market Prediction Based on Generative Adversarial Network", Procedia computer science, Vol. 147, pp. 400-406, 2019.
[4] Shi. L, Teng. Z, Wang. L, Zhang. Y and Binder. A, "DeepClue: Visual interpretation of text-based deep stock prediction", IEEE Transactions on Knowledge and Data Engineering, Vol. 31(6), pp.1094-1108, 2018.
[5] Chen. L, Qiao. Z, Wang. M, Wang. C. Du. R, and Stanley. H. E, "Which artificial intelligence algorithm better predicts the Chinese stock market?", IEEE Access, Vol. 6, pp. 48625-48633, 2018.
[6] Zhang. X, Qu. S, Huang. J, Fang. B, and Yu. P, "Stock market prediction via multi-source multiple instance learning", IEEE Access, Vol. 6, pp. 50720-50728, 2018.
[7] Long. W., Lu. Z and Cui. L, "Deep learning-based feature engineering for stock price movement prediction", Knowledge-Based Systems, 164, pp.163-173, 2019.
[8] Chen. Y and Hao. Y, "A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction", Expert Systems with Applications, Vol. 80, pp.340-355, 2017.
[9] Patel. J, Shah. S, Thakkar. P and Kotecha. K, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques", Expert Systems with Applications, Vol. 42(1), pp.259-268, 2015.
[10] Idrees. S. M, Alam. M. A and Agarwal. P, "A Prediction Approach for Stock Market Volatility Based on Time Series Data", IEEE Access, Vol. 7, pp. 17287-17298, 2019.
[11] Wen M, Li P, Zhang L, Chen Y, "Stock Market Trend Prediction Using High-Order Information of Time Series", IEEE Access, Vol. 7, pp. 28299-308, Feb 2019.
[12] Liu G, Wang X., "A numerical-based attention method for stock market prediction with dual information", IEEE Access, Vol. 7, pp. 7357-67, Dec 2018.
[13] Minh D. L, Sadeghi-Niaraki A, Huy H. D., Min K, Moon H., "Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network", IEEE Access, Vol. 6, pp. 55392-404, Sept 2018.
[14] Ren R, Wu D. D, Liu T, "Forecasting stock market movement direction using sentiment analysis and support vector machine", IEEE Systems Journal, Vol. 13(1), pp. 760-70, Mar 2018.
[15] Stak overflow, "http://stackoverflow.com/questions/19790188/"
[16] Moneycontrol, "https://www.moneycontrol.com/".
[17] Oliphant T. E., "A guide to NumPy", Trelgol Publishing USA, Vol. 1, 2006.
[18] McKinney, W., & others., "Data structures for statistical computing in python", In Proceedings of the 9th Python in Science Conference, Vol. 445, pp. 51–56, 2010.
[19] Bird, Steven, Ewan Klein, and Edward Loper, "Natural Language Processing with Python", O'Reilly Media, 2009.
[20] Chollet, Francois & others, "Keras", https://keras.io, 2015.
[21] Hunter J. D., "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.
[22] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., & others., "Scikit-learn: Machine learning in Python", Journal of Machine Learning Research, pp. 2825–2830, Oct 2011.