Stock Market Movement Prediction Using News Analytics

Nitin Sakhare[1,*] , Divya Midhun[2], Dharmesh Dhabliya[3]

[1]Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, India
[1]Lincoln University College, Malaysia.

[2]Lincoln University College, Malaysia.
[3]Department of Information Technology, Vishwakarma Institute of Information Technology, Pune, India.

nitinsakhare4@gmail.com, divya@lincoln.edu.my, dharmeshdhabliya@gmail.com

Abstract: The price of a stock is very erratic and predicting it accurately has always been awarded lucratively. One such method of prediction is achieved through deploying machine learning algorithms to speculate the direction of movement of a stock. Two main factors affect the price of a stock- the news around the sector and company, and the company financials. Both these aforementioned factors can be studied and used to predict a company's stock price and hence have become subject to extensive studies and exploration. Method: In the proposed decision support system, we have employed natural language processing, neural networks, and supervised machine learning algorithms to analyze the news about the sector, company, and country which results in a unique ensemble method to generate a 5-bit pattern, which is then mapped to 5 calls that range from a high chance of increase to a high chance of decrease. This ensemble approach eliminates the drawbacks of using these techniques exclusively and assembles the advantages in one system. Calls generated by this system can be used to aid our decisions while investing in the stock market, as we can construe the direction of the movement of the market.

Keywords: Sentiment Analysis; Stock Market; Decision Support System; Neural Networks; Ensemble Approach; Machine Learning

Introduction

The stock market does not solely rely on the fundamental analysis of a company; it is largely influenced by ongoing mundane affairs and the general sentiment of the public. The very idea of a stock market crash or a rally is a result of extremely varied and opposite sentiments of the investors. This was very well showcased by the worldwide stock market crashes in March 2020, on the onset of the Corona Virus Pandemic, and subsequent lockdowns that were imposed globally. The markets rallied several times when news and speculation related to vaccines, medicines, and a people-friendly budget was announced (Indian budget 2021). The general sentiment of the market can be derived through benchmark-based indices of stock exchanges of that country. India has two such indices that are widely used to exhibit the general inclination of the market, namely NIFTY 50 (National Stock Exchange) and SENSEX (Bombay Stock Exchange). These indices take into consideration the biggest companies (market-cap wise) listed on the exchange and use a weighted average system to calculate a number that is used to track the behavior of some famous large-cap companies, and in general, the Indian market.

Financial news channels, blogs, websites, forums, and print media use an established format of reporting news related to the stock market by using the words NIFTY AND SENSEX. In the proposed approach, we have scrapped news that has the words NIFTY and SENSEX using a Python library named BeautifulSoup. We have used the Economic Times [3] as our news outlet, as it has a fixed format of reporting news related to the Indian stock markets (NSE and BSE), which makes it easier to extract, transform, and load the data to feed into our model.

In the proposed system, based on our literature survey and study of related work, we have used an artificial neural network (ANN), Linear Support Vector Classifier (Linear SVC), Random Forest Classifier (RFC), Logistic Regression, and

a natural language processing technique named Dictionary Based Approach. These five algorithms correspond to 5 bits, in order of most significance to least significance, which has been considered while generating calls on a 5-point scale, starting from strong chances of increase, might increase, no considerable change might decrease, and high chances of decrease. It is a well-known and accepted fact that it is important to understand the direction of the movement of stock price more than the magnitude of its movement, which we have aimed to predict using the proposed system. Each model gives a binary output after classifying news as positive or negative.

By using natural language processing (NLP), neural networks, and supervised machine learning classification algorithms, we have tried to reduce the restrictions generated by using only one of the aforementioned approaches, which results in a final output based on an ensemble and novel technique. [1] used a beautiful soup library in Python to extract live data from web pages, a popular web-scraping tool along with selenium and scrapy. They have fetched live data of SENSEX & NIFTY and applied sentiment analysis to it. [2] used a dictionary-based approach and Simple Moving Average (a technical indicator) to tag the news as positive, neutral, and negative. [4] used Twitter API to collect data related to the company whose stock price is being predicted. Then sentiment analysis and LSTM were used to predict that company's stock price. [5] used bigrams and trigrams to understand the polarity of individual words based on the context of data scraped from MoneyControl and generated three calls- buy, sell, and hold. Sun et al. (2018) use Gaba, a Chinese financial stock forum to which a sentiment lexicon, GubaLex, is directed to collect basic sentiment lexicon. The data is then collected, preprocessed, filtered, and run through a trained PNN model to get the bullishness score based on the Sentiment words, Degree words, Exclamation Marks, and Negative Words. [7] proposed a theoretical model for a multi-agent decision support system for predicting sentiment from the data based on social media. The agents of the work cycle used were Data search, Data collection, Data Processing, Text Mining, Modelling, and Decision-making agents to predict the movement of the market and volatility with some intervals. [8] uses data from Sina Weibo and the Financial community and pre-processes it. They use a Linear Regression model and a Multilayer Perceptron Model to provide a strong relationship between social media data and the Chinese Composite Index to predict the stock price of the composite index. Li et al. (2018) used Twitter search API to retrieve tweets. They have used Yahoo Finance API for getting the historical stock data. Naive Bayes and Support Vector Machine are the two machine learning models that have been used to predict tweet sentiments and gain insight into the correlation between Twitter sentiment and stock prices. Simon Xie [9] gathered share price data using the Alpha Vantage platform. They collected Twitter data and used Latent Dirichlet Allocation (LDA) for topic modeling. Then, Sentiment Analysis was performed based on topics to find the correlation between the Twitter data and the share prices. Makrehchi [10] retrieved data from various social media platforms (for example, tweets from Twitter). Then, each tweet was labeled positive or negative. The model was trained on this collected data to predict labels for future tweets. Then, the net sentiment per day was calculated to show its correlation with the stock market movement. Sagala uses algorithms such as Support Vector Machine (SVM), Kth Nearest Neighbour (KNN), and Naive Bayes, to perform technical and sentimental analysis, and concluded that SVM gives the highest accuracy in predicting the stock price on a 5-trading day window[11].

Methodology

In our model, we have formulated a scraper script by bringing into play the Python package Beautiful Soup. The scraper model acquires the latest headlines from The Economic Times [3] for the NIFTY 50 index. The retrieved data is transfigured by pre-processing steps namely removal of punctuation marks, unwanted characters, and spaces. We have amalgamated the “Loughran McDonald” sentiment dictionary with “Henry’s financial dictionary” to concoct a thorough list of positive and negative words. We loop through each headline to analogize with the coalesced dictionary to enumerate the connotation of each headline. The dataset comprises a headline column and a sentiment column. The dataset contains 3572 records, which were balanced using a Python library named SMOTE. The labels for the dataset were calibrated manually as a dictionary-based approach is an unsupervised mode, and a well-labeled dataset is paramount for achieving high accuracy while training our supervised machine learning models and our

artificial neural network. The proposed system aims to provide an ensemble approach of neural networks, natural language processing, and supervised machine learning models. Our system is divided into five major modules, starting from importing files, extracting news from The Economic Times website [3], and pre-processing the data using the Natural Language Toolkit (NLTK) library in Python, after which the entire model splits into three different methods and 5 different algorithms. The next module consists of assembling the outputs from each of the models and generating a bit pattern. The final part translates this bit pattern into a call, on a scale of high chances of increase of stock price too high chances of decrease of the stock price.

1.1 Module 1 Importing files

We have used pre-trained supervised machine learning models for consistent accuracy, which need to be uploaded into the interactive Python notebook (ipynb) before starting the execution of the code. We linked Google Drive to our Google Colab notebook so that we do not have to perform the task of manually uploading the required files every time we start the runtime.

1.2 Module 2 Extracting news and performing pre-processing operations

We first navigate to the section of the Economic Times website [3] (https://economictimes.indiatimes.com/markets/stocks/news) where we get news related to NIFTY 50 and SENSEX. This URL is then added to our web scraper script which is designed using the BeautifulSoup library. A lot of undesirable data is obtained along with news headlines which we require. Our script contains a filter to remove those, which results in the required output of news headlines and time stamps. Economic Times has an undeviating format of publishing news related to NIFTY and SENSEX, which makes the headlines sufficient to be used as data to predict the movement. After extracting news from the website, we then perform some standard pre-processing operations on the data collected using the natural language toolkit library (NLTK), which makes it suitable to be count vectorized or to be directly fed into models. Some of these operations are transforming the data into lowercase letters, removing all stopwords (except 'not') as they do not alter the sentiment of data, and retrieving the stem of words using Porter Stemmer.

1.3 Module 3 Algorithms

This module consists of five models: artificial neural network (ANN), linear support vector classifier (linear SVC), Logistic Regression, Random Forest Classifier (RFC), and dictionary-based approach. They have been explained in detail in sub-sections 2.3.1 to 2.3.5.

1.3.1 Artificial Neural Network

Our Artificial Neural Network model is further split into two parts, training, and prediction.

Training ANN Model

The training process of the ANN model begins by creating a collection of the training dataset and performing pre-processing operations. This collection is then used as input for the count vectorizer component. The resulting count vectorized data is employed to train our ANN model. Our ANN model consists of a total of 4 layers, including an input layer and 2 hidden layers that utilize the RELU activation function. Each layer comprises nineteen neurons. The output layer uses the sigmoid activation function to classify data as either positive or negative. During compilation, we optimize the model using the 'Adam' optimizer and calculate the loss function using binary cross-entropy. To evaluate the performance of our model, we feed it the test set in batches of 32, iterating over 10 epochs to minimize the risk of overfitting.

Prediction using the ANN model

For the prediction part, the extracted news is first pre-processed and the corpus is formed, after which it is passed to the count vectorizer. The output of the count vectorizer is then fed to the trained ANN model, which then gives the prediction in the form of 1 or 0, with 1 being classified as positive and 0 being classified as negative.

1.3.2 Linear Support Vector Classifier (LSVC)

"Support Vector Machine" is a supervised machine learning algorithm, that performs classification by tracking down the hyper-plane that sets apart the classes. The linear SVM kernel deduces the hyperplane and reworks it to a linear problem.

Equation of Linear SVC kernel:

$$K(x, y) = x.y \quad (1)$$

The vectorized headline(x) is given as input to the linear support vector classifier model along the polarity of the headline is provided as the label (y).

1.3.3 Logistic Regression

Logistic regression is a linear classifier which uses a linear function

$$f(x)= b_0 + x_1b_1 + x_2b_2+ \dots + x_nb_n \quad (2)$$

where b1, b2… bn are the coefficients of the line of the perfect fit.

1.3.4 Random Forest Classifier (RFC)

Random forest is a supervised learning algorithm, an ensemble of decision trees, that tasks the bagging method. The bagging method consolidates learning models to improve the accuracy and stability of the model. The random forest algorithm utilizes the decision tree algorithm as the base learner alongside the bagging technique for sampling rows and columns.

1.3.5 Dictionary Model

The model receives the news scraped from The Economics Times one by one [3]. To process the news headline, a technique called word tokenization is applied, which involves tasks like converting all words to lowercase, removing unnecessary words, and classifying the words. The classification process is straightforward; every word in the headline is compared to the words in positive and negative dictionaries. However, if a word is found in either dictionary, it gets the label of that particular dictionary. To know the sentiment of the news, most of the labels in the sentence are analyzed. A negative label for most of the words classifies the sentence as a negative one, and vice versa.

For instance, consider this piece of news:

Sensex plunges (-1, neg) 883 points, Nifty50 below (-1, neg) 14,400; ONGC tanks 4%

Score = -2: Predicted sentiment = negative

1.4. Module 4 Bit Patterns Assembly

The idea of bit pattern assembly is derived from the work done by Joshi and Sakhare[13]. Our models predict the final class of the news as positive and negative and return 1 or 0 respectively. All five of the models give one bit each which are arranged in a binary fashion of most significant bit to least significant bit. For the model, ANN is the most significant bit followed by Linear Support Vector Classifier, Logistic Regression, Random Forest Classifier, and Dictionary being the least significant bit of all. Furthermore, as we get five bits of the total number of combinations, we get 32 possible combinations, ranging from 11111 to 00000.

1.5. Module 5 Calls

After getting the generated bit pattern, we map them to five calls, which are: High chance of increase, might increase, no significant change, might decrease, and a high chance of decrease.

Results and Discussion

Our results are presented in a comprehensible table format, which includes printing the news, the date and time, the individual models and their output, and the call generated. This way, the user can read the news and confirm his decision. Table 1 resembles the actual output we present. Table 2 is for validation, where we have achieved 70% accuracy with our calls for specific news we received. News related to stock market indices is often published before the markets open for trade and after they close. For this reason, we have taken the next trading day's price after the

market is open (not exactly the opening price, but the price after 30-60 minutes of the market opening for the public). This is done to consider the effect of the general and multiple news received, and as NIFTY 50 is an index, its price depends on the movement and volatility of the 50 constituent companies.

Table 1. Results Table with 5-bit pattern and final call output.

News	ANN	LSVC	LR	RFC	DBC	Calls
"Market Movers: What's fueling Asian Paints; HUL's race to catch Infosys; EITC's good day"	1	1	1	1	1	High chance of an increase
'Sensex gains 42 points, Nifty ends below 14,700; Asian Paints soars 8%"	1	1	1	1	1	High chance of an increase
"Sensex struggles near flatline as rising US inflation spooks investors: What else is impacting D-Street"	1	1	1	1	1	High chance of an increase
"Stock market update: 94 stocks hit 52-week highs on NSE"	1	1	1	1	0	High chance of an increase
"Dalal Street indices fall 1% on Asian cues"	0	1	1	1	0	No significant change
"Economic normalization likely to return by Aug-Sept: CLSA"	0	0	0	1	0	High chance of a decrease
"Nifty can crash up to 1,000 pts, next one week crucial: Analysts"	0	1	1	1	0	Might decrease
"Sensex plunges 883 pointsNifty50 below 14,400"	0	0	0	0	0	No significant change
"Sensex extends losses to 2nd day amid rising infections, US inflation worries"	1	0	1	1	0	Might increase
"Sensex rallies over 600 pts, Nifty reclaims 15K; RIL, metal stocks shine"	1	1	1	1	0	High chance of an increase

Table 2. Validation table

News	Date	Model Call	Price at the exact time	Price after some time
"Market Movers: What's fueling Asian Paints; HUL's race to catch Infosys; ITC's good day"	14-05-2021 16:43	High chance of an increase	14,681.00	14,938.00

“Sensex gains 42 points, Nifty ends below 14,700; Asian Paints soars 8%”	14-05-2021 15:51	High chance of an increase	14,681.00	14,938.00
“Sensex struggles near flatline as rising US inflation spooks investors: What else is impacting D-Street”	14-05-2021 09:32	High chance of an increase	14,628.75	14,749.65
“Stock market update: 94 stocks hit 52-week highs on NSE”	14-05-2021 13:07	High chance of an increase	14,664.25	14,648.40
“Dalal Street indices fall 1% on Asian cues”	13-05-2021 08:14	significant change	14,696.50	14,749.40
“Economic normalization likely to return by Aug-Sept: CLSA”	14-05-2021 07:08	High chance of a decrease	14,696.50	14,749.40
“Nifty can crash up to 1,000 pts, next one week crucial: Analysts”	19-04-2021 17:00	Might decrease	14,359.45	14,526.95
“Sensex plunges 883 pointsNifty50 below 14,400”	19-04-2021 08:00	significant change	14,617.85	14,306.60
“Sensex extends losses to 2nd day amid rising infections, US inflation worries”	12-05-2021 15:38	Might increase	14,696.50	14,749.65
“Sensex rallies over 600 pts, Nifty reclaims 15K; RIL, metal stocks shine”	29-04-2021 22:20	High chance of an increase	14,894.90	14,855.45

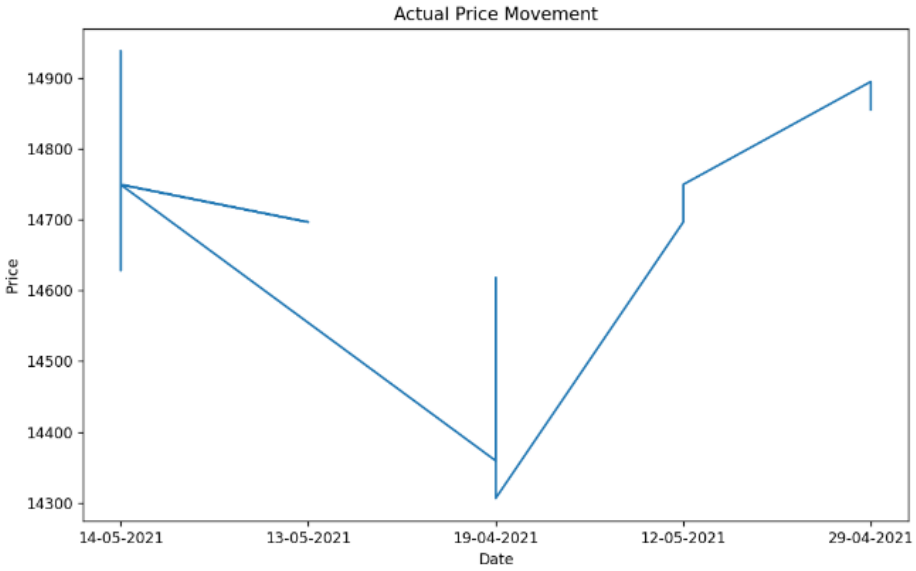

Figure 1. *A line graph showing the actual price movement over time.*

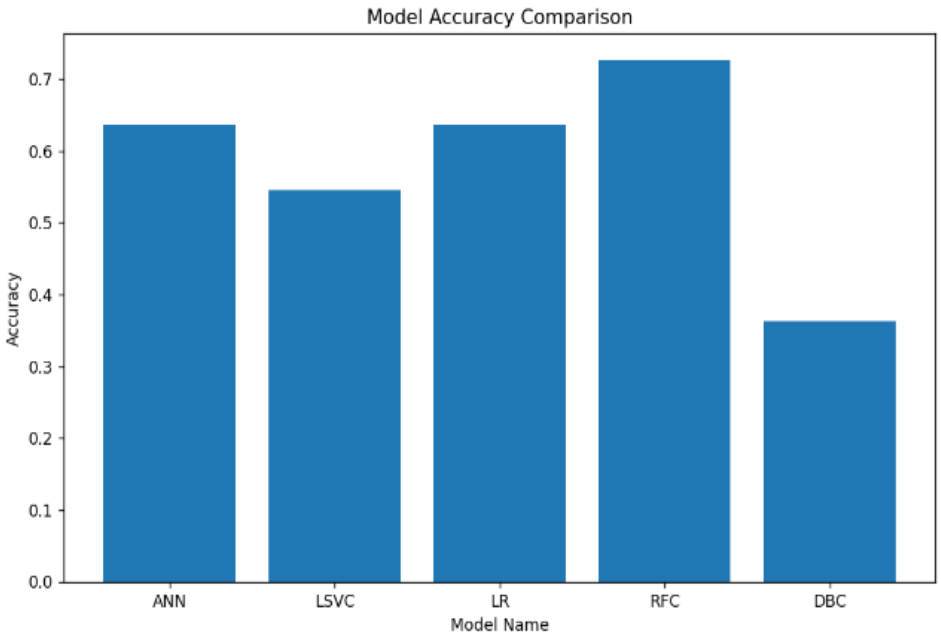

***Figure 2**. A bar graph comparing the accuracy of different machine learning models.*

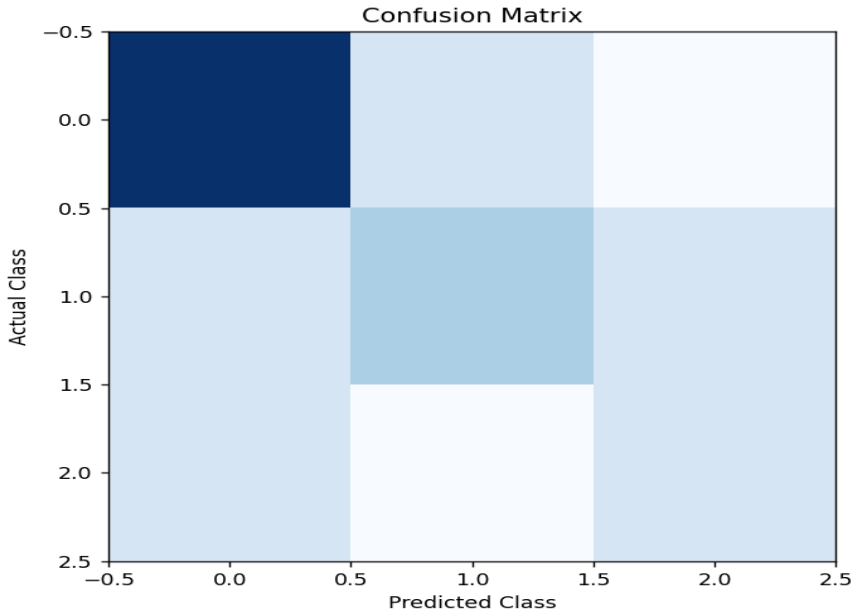

***Figure 3.** A heatmap representing the confusion matrix.*

Discussion

The results of our study show that sentiment analysis can use a combination of machine learning techniques to predict the stock market. Predictions of our proposed system based on combining the predictions made by multiple machine learning models are more accurate than other individual models. The following bar graph proves that the

accuracy achieved by the proposed system is 70% and higher in comparison to other individual models. The line graph reveals that in most cases, the actual price movement follows the predicted trend.

Our study is more accurate than the studies available in the literature that rely on sentiment analysis and machine learning for predicting stock market trends. For instance, earlier studies achieved an accuracy between 60% and 65%, while our proposed system can achieve an accuracy of 70%. This implies that our approach is more effective for predicting stock market trends.

The results of our study have implications for investors and financial analysts who can use our proposed system to make informed investment decisions. Our system can provide early warnings of potential market trends by analyzing the sentiment of financial news articles, allowing investors to make more informed decisions. Our system can also be used to monitor market sentiment in real time, providing valuable insights for investors and financial analysts.

One limitation of our study is that it uses a limited dataset and may not generalize well to other datasets.

Future Work

Our system can further be enhanced using additional data sources including social media data from Twitter, Facebook, and news articles or other text sources. Other types of alternative data sources include satellite imagery and sensor data, through which the predictability of our system can further be improved.

Another possible direction for future work is the development of enhanced machine learning models capable of better accurate predictions. This can be done by using deep learning techniques like CNN and RNN or employing ensemble methods including bagging and boosting. Transfer learning is also applied by leveraging existing pre-trained models for further improvements in the system.

Finally, future work can also focus on improving the explainability and transparency of our system. This can be achieved by providing more informative explanations of its predictions, utilizing visualization techniques to facilitate understanding of the decision-making process, and developing methods to detect and prevent bias in its predictions.

Conclusion

The proposed decision support system uses an ensemble approach of natural language processing, supervised machine learning, and artificial neural networks. It can be used to make the final judgment in investing (index fund) and understanding the general sentiment of the public to forecast the direction of change of NIFTY, and the five calls will increase the confidence of the investor by giving them a much more precise insight towards making the decision. However, if there is a possibility of extracting enough data to train the models, and to get access to periodical news updates about that company, it can be added to the model, for people looking to assess specific costs. In this system, we have predicted the direction of change, which is often the best one can do in a system as volatile as the Indian Stock Markets (NSE and BSE), yet if we can figure out a way to employ a Recurrent Neural Network (RNN) to predict the magnitude of change, it would be a great addition to this decision support system.

References:

1. Bhardwaj A, Narayan Y, Dutta M. Sentiment analysis for Indian stock market prediction using Sensex and nifty. Procedia Computer Science. 2015 Jan 1;70:85-91. https://doi.org/10.1016/j.procs.2015.10.043.
2. Bharathi S, Geetha A. Sentiment analysis for effective stock market prediction. International Journal of Intelligent Engineering and Systems. 2017;10(3):146-54. https://doi.org/10.22266/ijies2017.0630.16
3. The Economic Times
4. Panday H, Vijayarajan V, Mahendran A, Krishnamoorthy A, Prasath VB. Stock Prediction using Sentiment Analysis and Long Short-Term Memory. European Journal of Molecular & Clinical Medicine. 2020 Nov 28;7(2):5060-9.
5. Shah D, Isah H, Zulkernine F. Predicting the effects of news sentiments on the stock market. In2018 IEEE International Conference on Big Data (Big Data) 2018 Dec 10 (pp. 4705-4708). IEEE. https://doi.org/10.1109/BigData.2018.8621884

6. Sun Y, Fang M, Wang X. A novel stock recommendation system using Guba sentiment analysis. Personal and Ubiquitous Computing. 2018 Jun;22(3):575-87. https://doi.org/10.1007/s00779-018-1121-x
7. Chornous G, Iarmolenko I. Decision support system for predicting stock prices based on sentiments in social media. In Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing 2017 Mar 22 (pp. 1-4). https://doi.org/10.1145/3018896.3025158
8. Chen G, He L, Papangelis K. Sentimental Analysis of Chinese New Social Media for stock market information. In Proceedings of the 2019 International Conference on Pattern Recognition and Artificial Intelligence 2019 Aug 26 (pp. 1-6). https://doi.org/10.1145/3357777.3357778
9. Li M, Yang C, Zhang J, Puthal D, Luo Y, Li J. Stock market analysis using social networks. In Proceedings of the Australasian Computer Science Week Multiconference 2018 Jan 29 (pp. 1-10). https://doi.org/10.1145/3167918.3167967
10. Xie S, Li M, Li J. Sentiment Correlation Discovery From Social Media to Share Market. In Proceedings of the Australasian Computer Science Week Multiconference 2019 Jan 29 (pp. 1-8). https://doi.org/10.1145/3290688.3290712
11. Makrehchi M, Shah S, Liao W. Stock prediction using event-based sentiment analysis. In2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) 2013 Nov 17 (Vol. 1, pp. 337-342). IEEE. https://doi.org/10.1109/WI-IAT.2013.48
12. Sagala TW, Saputri MS, Mahendra R, Budi I. Stock Price Movement Prediction Using Technical Analysis and Sentiment Analysis. In Proceedings of the 2020 2nd Asia Pacific Information Technology Conference 2020 Jan 17 (pp. 123-127). https://doi.org/10.1145/3379310.3381045
13. Joshi S, Sakhare N. History Bits based novel algorithm for classification of structured data. In 2015 IEEE International Advance Computing Conference (IACC) 2015 Jun 12 (pp. 609-612). IEEE. https://doi.org/10.1109/IADCC.2015.7154779