Sentiment Analysis of Stocks Based on News Headlines Using NLP

Aastha Saxena, Arpit Jain, Prateek Sharma, Sparsh Singla, and Amrita Ticku(✉)

Computer Science and Engineering Department, Bharati Vidyapeeth College of Engineering, New Delhi, India
amritaticku27@gmail.com

Abstract. In todays' world everyone starting from a child to an adult is studying stocks and is finding ways to earn more by studying the patterns of the market. Stock market is a compound interrelated system of various investors. It fluctuates frequently and hence is hard to predict what is yet to come. All the companies worldwide rely on these forecasts and speculations so that they can increase their profits. Everyday news on the economic front plays a vital role in defining the jump or drop in the prices of stocks. The market news helps the investor excessively in determining his bidding as it is immensely rich in information. In this study, we extract useful information from news headlines of a particular company to investigate the immediate impact of it on the company's stock growth. Having the text based dataset we use NLP and compare two approaches using two different algorithms which both collectively determine the sentiment of the news headline (whether it is positive/negative/neutral) in lieu with the company stocks.

Keywords: Natural Learning Process · Convolution Neural Network · Gated Recurrent units

1 Introduction

Quantitative traders with lots of money buy stocks derivatives from stock markets and equities at cheap rates and then sell them at higher costs later. This prediction fashion is not new in the stock market and still it is the topic of debate in many organizations. Majorly there are two ways to analyze the market which investors practice before putting their money in the market. The first one being the analysis of fundamentals and the other being the technical analysis. In the fundamental analysis, investors focus on the innate stock values, country economy, political scenarios occurring at the point of time, industrial conducts etc. to decide whether to invest or not. While technical analysis is to study the progressment of the stocks by understanding the statistical measurements generated by the market activity, which includes stock volumes and former stock values. Due to the soaring influence of machine learning in copious applications and domains, it has led the investors and traders to inculcate these techniques of machine learning for their own discipline of work, and most of those have produced promising results.

V. E. Balas et al. (Eds.): AITEES 2022, AHIS 3, pp. 124–135, 2023.
https://doi.org/10.2991/978-94-6463-074-9_12

This market of stocks in the main follows the path of randomness, as in we can get the finest predictions about the rates of stock prices to come by looking at their today's price value. Forecasting stock indices is certainly a tedious task due to the unstable nature of this market that demands a high accuracy predicting model. The stock market prices are very waverly which affects the investor's faith in models. The nature of stock prices is quite dynamic and changes frequently because of the complex nature of the financial domain. The stock market predictions are the probable future stock prices, in this type of predictions there are chiefly two types which are real time predictions and dummy predictions which are used in the stock market prediction system. In dummy predictions there are a determined set of rules which foretell the future price by calculating the average price. Whereas in the real time prediction, they use the internet and see the current price of shares of the company.

Researches based on sentiment analysis are quite common, moreover modern technologies like neural networks are also being used by the researchers to have an insight about the real time data of stock movements. The study 'Optimization of sentiment analysis using machine learning classifiers' [1] has used various intricate machine learning algorithms like Naive Bayes, J48, OneR and BFTree for improving the results of sentiment analysis. Through the research it is discovered that Naïve Bayes is quite quick in models whereas OneR is found to be more budding in generating the precision accuracy of about 91.3% and 92.34% in classified instances which is some concrete and promising information for our study. Also, deep learning techniques are being employed for the rational study of stocks which have been unpredictable for a long time. Research study - 'Stock movement prediction with sentiment analysis based on deep learning networks' [2] which uses CNN, GRU, LR as classifiers, PCA algorithm for their findings have been quite useful for our study. PCA confirms to be an keen and efficient feature decomposition method in comparison to all the other algorithms. Also, the results in their study showed that LR obtained the most stable performance while the CNN and GRU had the highest accuracy among the classifiers.

The study of predicting the behavior of stock prices in the coming future i.e. whether the stock prices surge or take a hit, takes into account the financial news available in the form of data on Kaggle platform. The first task which comes in the picture is to convert the long paragraphs of news as our dataset into a form which can be feeded to a deep learning model. Different language preprocessing techniques are used to convert data into machine interpretable form. The BERT-base-uncased model is used as the base model in our research. As observed from previous research, sentiment analysis with BERT has come out to be one of the most efficient tools in the field. The model helps in understanding the effect of customized BERT neural networks in the forthcoming stock predictions. It also highlights issues that are being faced by following this approach, that will be essential for future research work in this area.

2 Literature Survey

Following are some brief summarized work in chronological order, previously performed by peoples/groups/organizations in our concerned field of sentiment analysis. In 2018, Mukhtar et al. [3] use emotional analysis on Urdu blogs that they have acquired in a few

domains using the Guided Reading Machine and Lexicon-based models. For Lexicon-based models, accurate Urdu sentiment analysis and various Urdu Sentiment Lexicons were used, and a Supervised Machine learning (SVM) algorithm, DT, KNN, and SVM were used. Data will be collected from two sources for the best possible emotional analysis. After obtaining the results, they saw lexicon-based models surpassed models based on machine-readable readings.

In 2018, Abdi et al. [4] provide a machine-based strategy to summarize the views of users mentioned in the reviews. This approach combined many types of features into a unique feature set for modeling precise differentiating models. From now on, in order to determine the best performance, a feasibility study is conducted to select the four best feature models and seven class dividers to select the appropriate feature set and create an effective machine learning algorithm. The method they followed was already implemented using various data sets. The results and test results explained that the combination of IG as a feature selection factor and SVM-based differentiation method has improved performance in many ways.

In 2019, Ray and Chakrabarti [5] they have introduced an algorithm based entirely on in-depth reading concepts to extract features in the text and analysis of user emotions about the feature. In all of his sentence-based emotional expressions, seven layers of Deep CNN were used to tag the features. In order to improve the effectiveness of emotional scoring and feature-releasing models, authors incorporate in-depth reading methods using a variety of rules-based models. Finally, it was noted that the proposed method achieved better accuracy compared to all other methods.

The research paper - *'Stock Market Prediction Using Machine Learning'* [6] based on stock market predictions using Vector support (SVM) and Radial Basis Function (RBF). In this study, they proposed the use of data collected from various global stock markets with the help of machine learning models and algorithms to predict the stock index movement. SVM did not present a problem of overcrowding of any kind. Various machine-based models have been suggested to predict the daily trends of Market stocks. The numerical results suggested higher efficiency. The model produced a higher profit compared to the selected benchmarks.

The research in data mining and neural networks techniques [7] has employed traditional methods that may not ensure the reliability of the prediction. In this paper, They basically go over the two techniques namely mining of data and the neural network in artificial intelligence. Since the neural network was able to extract important news from a large dataset and data mining can detect the trends of future models. However, a combination of both these techniques made it a lot more reliable. Many methods of predicting stocks used a lot of techniques that were totally based upon neural networks. After completing the research, it was clear that data mining and neural networks were very reliable techniques to deal with any type of unpredictable data like stock data that involves any kind of prediction.

This research 'Stocks Market Prediction using Support Vector Machines' [8], provides strong evidence that old predictive regression models had come across various challenges and the main reason behind that was uncertainty of models and instability of different parameters. They believed that their model still had a great amount of room for improvement. Improvement can be achieved by adding certain types of variables

to the model, mainly those showing aspects of the company which were not related to profitability or about earnings-share relationships. They believed that multiple statistics could be used to measure the value of the company with respect to global economic conditions and its own financial performance of last year.

After the introduction of Artificial Intelligence, coded functions of prediction had proved to be much more helpful in the prediction of stock prices rather than the old theories. In this Paper [9], Artificial Neural Network (ANN) and Random Forest techniques (RFS) had been applied on a dataset in order to get the best prediction of future stocks. Every type of financial data was used in prediction of the stock. The models were then evaluated using standard indicators: RMSE and MAPE. They have analyzed on the basis of RMSE and MAPE and they observed that an artificial neural network was much better than a random forest. After completion of research, Results showed that the best values obtained by the ANN model give RMSE (42%), MAPE (77%).

Stock prediction played a major role for the economic situation of our country which increased the curiosity of developers to make new models for their prediction. ARIMA models had been used in the literature for the time series prediction a lot of times in the past. This paper [10] brings a large-scale process of building the stock price prediction models using the ARIMA model. Dataset was taken from the Nigeria Stock Exchange (NSE) and New York Stock Exchange (NYSE) and was used for the price prediction of stocks using the desired models. Results revealed that, ARIMA model has a very strong potential for short-term predictions. This paper presented a comprehensive process for creating an ARIMA model for stock price prediction. Test results of this model can only be used for short-term stock forecasting. This may guide investors in the stock market to make more profitable and secure investment decisions. With the results provided, ARIMA models can compete effectively and efficiently to work with emerging forecasting strategies in the event of short-term forecasting.

The study short-term stock market price trend prediction using a comprehensive deep learning system [11] apply Fee Augmentation (FE) method with recursive feature elimination (RFE), followed by key component analysis (PCA), to develop or build a well-functioning and efficient feature engineering process. The system was customized by integrating feature engineering process with the LSTM predictive model, PCA had significantly improved the efficiency of the LSTM model training by 36.8%.

The paper - 'Predicting Stock Market Price Movement Using Sentiment Analysis' [12] use the MLP-ANN model in their research as Emotional analysis improved as predictive window sizes increased as it was possible for the MLP-ANN model to understand common sense in SNSs data sets with better accuracy, and to detect its effects on future stock price movements. It was also noted that the spread of news headlines had a better effect (64%) on stock volume than the number of comments made on the news (62.2%). The results obtained showed that the trading behavior of Ghanaian investors was partly. Influenced by social media.

The research - *'A novel ensemble deep learning model for stock prediction based on stock prices and news'* [13] uses an integrated reading model that combines LSTM and GRU. The model can reach up to 67% of MDA in the external capture test database. Lastly, Prediction models for the Indian stock market's research [14] utilizes various machine learning algorithms on the dataset which were Boosted Decision Tree, Logistic regression and SVM. It states that higher accuracy can be achieved using Boosted

decision tree algorithm i.e. 76.9%, as compared to logistic regression which gave an accuracy of 45% only.

3 Methodology

This section talks about the architecture of the model proposed in this paper. It also has the details about the dataset on which the model is trained and tested. It provides detailed information on BERT, dataset and preprocessing methods used, followed by the procedure used to create the model.

3.1 Dataset

For this research, the NewsHeadlines dataset is being used. The dataset is a combination of the stock price shifts and the world news present on Kaggle as - **"Daily News for Stock Market Prediction"**. There are 25 columns of the everyday news headlines in the data frame. The data ranges from around 2010 to 2020 along with the data from Yahoo finance ranges from 2000 to 2010 was collected. Labels are according to the Industrial Average stock index of Dow Jones. Class 1 → Increased price of stock. Class 0 → Decreased or same price of stock.

3.2 Preprocessing

Models cannot use raw text directly, so it is up to us to clean the text ourselves. Preprocessing of our financial news data is done through NLTK which stands for Natural Language Toolkit. NLTK is a powerful python suite of libraries and modules to carry out simple to complex natural language processing (NLP) on our data.

- **Converting data to lowercase and removing punctuations:** For the sake of simplicity, all the sentences of our data are converted to lowercase letters. It helps to maintain the consistent flow during the Natural Language Processing tasks and text mining. Punctuation removal is then performed on the data as punctuations create ambiguity while training the model and add to the noise in our data. Noise is referred to as the unnecessary elements present in our data that have no contribution in the training of the model, further increasing junk in our sentences.
  Example:
  Input: Really!! Zomato got hacked?
  Output: Really Zomato got hacked
- **Sentence tokenization and stop words removal:** Sentence Tokenization is the process of breaking down sentences into separate words. The *word_tokenize module* is used for this purpose, it breaks the sentences into tokens and these words act as an input for the further cleaning process. After this stop words are also removed from the data. Stop words are the most common words which are found in a language like conjunctions, articles, pronouns, etc. Some examples of stop words are 'a', 'the', 'and', 'or' etc. These words don't carry much information and can create an interference

with our original important words. Removing stop words helps the model to consider only key features for our model training.

Example:
Input: Will hacking affect the price of stock?
Output: hacking affect the price stock

- **Normalization:** It is an advanced step which is used to generate the grammatical root form of the inflected word to maintain uniformity. Stemming and Lemmatization are the two normalization processes that reduce the word to its common base form. The two processes though have the same task to perform but they are quite different from each other. **Stemming** - Stemming follows the crude strategy and essentially chops off letters from the front or end of the word until it's stem is reached. It takes into account a list of common prefixes and suffixes that can be found in an inflected word. For an example the stem for the words [talking, talked, talks] is talk. The process of stemming is faster than lemmatization but sometimes no correct stem is formed of the given word like for the word fought, the root word fight cannot be obtained through stemming.

  Example:
  boat → boat
  boating → boat
  boater → boat
- **Lemmatization** - In the process of Lemmatization word reduction is performed and a language's full vocabulary is considered to perform morphological analysis to the words in hand. As this process is based on linguistics and performing intelligent operations on the word, therefore it takes more time than stemming.

  Example:
  see → see
  saw → see
  seen → see
- **Creating a bag of words:** As most of the statistical algorithms, like machine learning or deep learning models, work with numeric data, therefore we need to convert our textual words into numbers, to be able to put the information gathered so far in a model. Some common approaches to performing this task are Bag of Words(BOW), TF-IDF, word2vec, etc. We use the bag of words technique which is a representation of text that describes the frequency of words within a document. We only keep into account the occurrence of words and disregard the word order or other grammatical details. The model is called so because any information about the order or structure of words in the document is discarded and it is only concerned with whether the known words occur in the document, not its location.

  Example:
  Review 1: Zomato is planning to open a new branch.
  Review 2: Zomato is hiring.
  OUTPUT: a branch hiring is new open planning to Zomato

3.3 BERT

Bidirectional Encoder Representations from Transformers, commonly known as BERT studies the contextual relationships between the words. The transformer has an encoder

which reads text as input and has a decoder which predicts. On the other hand, the directional model analyzes text in the sequential order. The encoder reads the complete text simultaneously.

There are many versions of pre-trained models for BERT available. Two of them are-

- BERT-base: 12 multi-head attention heads + 12 encoder layers + 768 hidden units: 110M parameters.
- BERT-large: 16 multi-head attention heads + 1024 hidden units + 24 encoder stack layers: 340M parameters.

Before using the pre-trained BERT model, input data should be converted to a proper format. It is done to obtain relevant embeddings for every sentence. Every encoder layer in the model takes a list of tokens as input and in output gives the equal embeddings with the same hidden size. For the classification task, the input vector must be given to the classifier, and "CLS" (the first token) of output is used by the classifier for classification.

4 Proposed Architecture

In this research, a BERT-base-uncased model was used which has a 768 hidden sized feed-forward network. This model takes input of IDs and the attention mask provided to the input. BERT Tokenizer uses the input sequences as to attach [CLS] and [SEP] joined together to the sentence at the start and the end respectively. The input ids and the attention masks will serve the output. These are passed to the BERT model and a vector of hidden size 768 for each token will come as an output.

To classify, the [CLS] is passed to the classifier. The classifier contains a 128 sized feed-forward linear layer, Batch Normalization layer, Dropout (0.6) and a final feed-forward layer of size 32, and output size 2 that classifies the news into real or fake. A 0.9 threshold was taken for the purpose of choosing one cell's output over the other. Figure 1 represents the proposed model architecture.

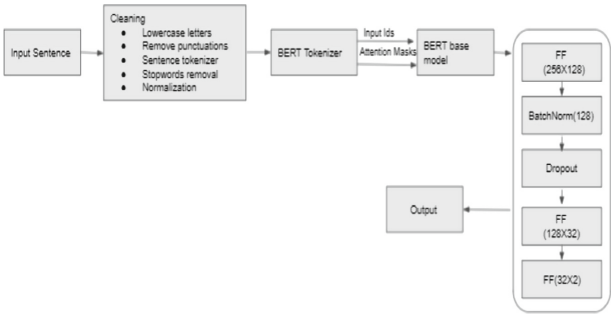

Fig. 1. Flowchart

5 Experimental Results

The experimental results that are obtained upon training the models on the given dataset are discussed and a comparison of the above-mentioned model with different models of classification and a pre-trained BERT model is made. Model training was executed on 80% of data points randomly selected. Testing of the models was done using the rest of 20% of the data.

Models like SVM and Random Forest Classifier were trained to compare the performance of novel classification approaches on the dataset. Furthermore, a vanilla BERT model was trained on the same parameters as the above-mentioned model, in order to measure the improvement made by the proposed model. All the baseline models were trained without any modification or fine-tuning.

The models were then compared on the basis of Accuracy (i), Precision (ii), Recall (iii), and F1 score (iv).

$$Accuracy = \frac{Number\ of\ samples\ classified\ correctly}{Total\ number\ of\ Samples} \tag{1}$$

$$Precision = \frac{True\ positives}{True\ positives + False\ postives} \tag{2}$$

$$Recall = \frac{True\ positives}{True\ positives + False\ Negatives} \tag{3}$$

$$F_1 Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

Comparing the pre-trained BERT model and the proposed model, and accuracy improvement of 5.60% was observed on the NewsHeadlines dataset (Table 1).

5.1 Calculation for SVM

$$Accuracy = \frac{3164}{4000} = 79.1\% \tag{5}$$

$$Precision = \frac{1640}{1640 + 471} = 77\% \tag{6}$$

$$Recall = \frac{1640}{1640 + 201} = 89\% \tag{7}$$

$$F_1 Score = 2 * \frac{0.77 * 0.89}{0.77 + 0.89} = 83\% \tag{8}$$

5.2 Calculation for Random Forest

$$Accuracy = \frac{3224}{4000} = 80.6\% \tag{9}$$

$$Precision = \frac{1668}{1668 + 553} = 75\% \tag{10}$$

Table 1. Training

Epochs	Training Acc	Training Loss	Val Loss	Val Acc	F1 Score	Precision	Recall
1	59.80	0.697	0.543	77.50	0.80	0.87	0.75
2	71.99	0.557	0.435	77.68	0.79	0.90	0.71
3	80.83	0.409	0.362	82.95	0.85	0.92	0.79
4	84.45	0.362	0.390	81.79	0.83	0.94	0.74
5	87.69	0.304	0.331	84.02	0.86	0.91	0.81
6	89.74	0.276	0.330	84.20	0.86	0.91	0.82
7	92.55	0.208	0.333	84.02	0.86	0.89	0.84
8	95.56	0.162	0.354	84.46	0.87	0.90	0.83
9	95.99	0.133	0.427	84.64	0.86	0.95	0.79
10	96.76	0.110	0.415	84.64	0.86	0.93	0.81
11	96.88	0.098	0.380	85.36	0.88	0.88	0.87
12	97.92	0.076	0.380	85.62	0.88	0.87	0.90
13	98.96	0.051	0.387	86.25	0.88	0.90	0.87
14	99.00	0.047	0.436	86.52	0.88	0.94	0.84
15	99.42	0.039	0.428	86.25	0.88	0.89	0.88

$$Recall = \frac{1640}{1640 + 383} = 81\% \tag{11}$$

$$F_1 Score = 2 * \frac{0.75 * 0.81}{0.75 + 0.81} = 78\% \tag{12}$$

5.3 Calculation for BERT

$$Accuracy = \frac{3450}{4000} = 86.25\% \tag{13}$$

$$Precision = \frac{1726}{1726 + 190} = 90\% \tag{14}$$

$$Recall = \frac{1726}{1726 + 261} = 87\% \tag{15}$$

$$F_1 Score = 2 * \frac{0.90 * 0.87}{0.90 + 0.87} = 88\% \tag{16}$$

See Table 2 and Figs. 2, 3 and 4.

Table 2. Comparison of all algorithms/methods

Model	Accuracy (%)	Precision	Recall	F1 Score
SVM	79.10	0.77	0.89	0.83
Random Forest	80.60	0.75	0.81	0.78
BERT	86.25	0.90	0.87	0.88

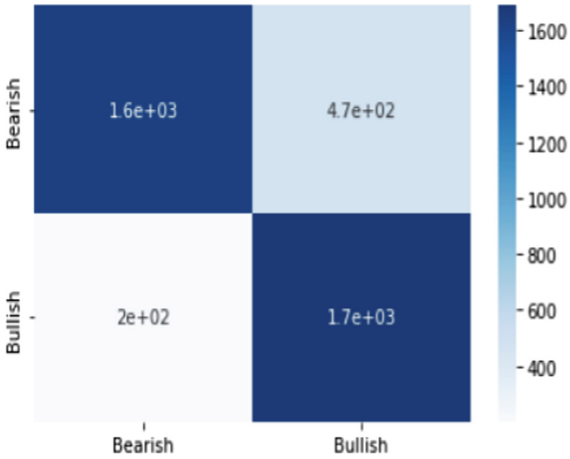

Fig. 2. SVM Confusion Matrix

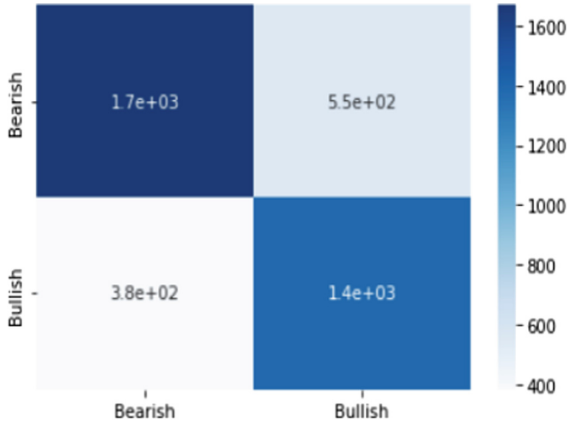

Fig. 3. Random Forest Confusion matrix

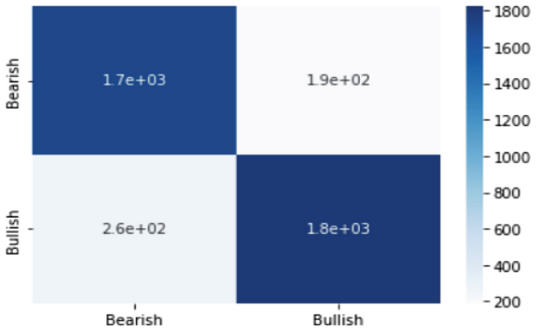

Fig. 4. BERT Confusion Matrix

6 Conclusions

In this paper, we proposed a method to predict the behavior of stock prices in future. i.e. either bearish or bullish using BERT based approach, we used a dataset from kaggle namely Daily News for Stock Market Prediction containing around 4116 images. We used 2 other algorithms i.e. SVM and Random forest classification. After comparison of all the three models, we achieved the best results using the BERT model with an accuracy of **86.25%**.

There are some issues when dealing with emotional keywords from tweets with multiple keywords. It is also difficult to manage the mispronunciation of words and phrases. Processing in many ways is limited to extracting foreign words, thumbnails and long words and their proper sense. Emotional analytics work has the potential to improve pre-word processing through deep neural networks and can extend this research to neural convolution networks.

References

1. Singh, J., Singh, G., Singh, R.: Optimization of sentiment analysis using machine learning classifiers. HCIS **7**(1), 1–12 (2017). https://doi.org/10.1186/s13673-017-0116-3
2. Shi, Y., Zheng, Y., Guo, K., Ren, X.: Stock movement prediction with sentiment analysis based on deep learning networks. Concurr. Comput. Pract. Exp. (2020). https://doi.org/10.1002/cpe.6076
3. Mukhtar, N., Khan, M.A., Chiragh, N.: Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains. Telemat. Inform. **35**(8), 2173–2183 (2018)
4. Abdi, A., Shamsuddin, S.M., Hasan, S., Piran, J.: Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment. Expert Syst. Appl. **109**, 66–85 (2018)
5. Ray, P., Chakrabarti, A.: A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. Appl. Comput. Inform. (2019)
6. Vanukuru, K.: Stock Market Prediction Using Machine Learning (2018). https://doi.org/10.13140/RG.2.2.12300.77448

7. Das, D., Shorif Uddin, M.: Data mining and neural network techniques in stock market prediction: a methodological review. Int. J. Artif. Intell. Appl. **4**(1), 117–127 (2013). https://doi.org/10.5121/ijaia.2013.4109
8. Hu, Z., Zhu, J., Tse, K.: [IEEE 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII) – Xi'an, China (2013.11.23–2013.11.24)] 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering - Stocks market prediction using Support Vector Machine, pp. 115–118 (2013). https://doi.org/10.1109/iciii.2013.6703096
9. Vijh, M., Chandola, D., Tikkiwal, V.A., Kumar, A.: Stock closing price prediction using machine learning techniques. Procedia Comput. Sci. **167**, 599–606 (2020)
10. Farzi, R., Bolandi, V.: Estimation of organic facies using ensemble methods in comparison with conventional intelligent approaches: a case study of the South Pars Gas Field, Persian Gulf. Iran. Model. Earth Syst. Environ. **2**, 105 (2016)
11. Neethu, M.S., Rajasree, R.: Sentiment analysis in twitter using machine learning techniques. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (2013). https://doi.org/10.1109/icccnt.2013.6726818
12. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL 2004, Stroudsburg, PA, USA. Association for Computational Linguistics (2004)
13. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. **36**, 105–139 (1999)
14. Kaufmann, S.: CUBA: artificial conviviality and user-behaviour analysis in web-feeds. Ph.D. thesis, Universität Hamburg, Hamburg, Germany (1969)