Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Forecasting the Direction of Stock Trends Using Machine Learning and Twitter

Rahul Bhatt[1], Gesu Thakur[2], Luxmi Sapra[3]
[1]Assistant Professor, Computer Science & Engineering, School of Computer Science & Engineering, Dev Bhoomi Uttarakhand University, Chakrata Road, Manduwala, Naugaon, Uttarakhand 248007
[2, 3]Associate Professor, Computer Science & Engineering, School of Computer Science & Engineering, Dev Bhoomi Uttarakhand University, Chakrata Road, Manduwala, Naugaon, Uttarakhand 248007
[1]socse.rahul@dbuu.ac.in, [2]head.ca@dbuu.ac.in, [3]socse.luxmisapra@dbuu.ac.in

Article Info
Page Number: 2729 - 2738
Publication Issue:
Vol 71 No. 4 (2022)

Abstract
As a consequence of technical advancements and the creation of new machine learning models, there has been increased interest in the study of stock market data. This is because these models provide traders and businesspeople with a platform from which to choose more lucrative companies. The information obtained from sources that are considered to be conventional media has had a considerable impact on the movement of stock prices. However, in recent times, platforms for online social networks have been pushing for a strategy that is more efficient in the dissemination of this information. The information that can be found on social networks may be of great use in determining the various perspectives and sentiments that individuals have on certain topics. Because of the volume and variety of these data, an improved machine learning model is one of the options that is continually being investigated for use in daily forecasting. As a direct consequence of this, an exhaustive comparative examination of the stock market models that have previously been put into use has been carried out as part of this effort. The information on Apple stock comes from Yahoo Finance and Kaggle, respectively. The classification technique did not provide sufficient confidence to be used to connect stock movement with feelings expressed on Twitter, with accuracy values ranging from 53 to 56 percent. This prevented the approach from being deployed. Utilizing the regression tactic, as opposed to the classification approach, resulted in superior outcomes. These models mostly depended on previous prices, and as is evident from the patterns, the emotion on Twitter was not a good predictor of changes in stock values. The XG Boost regressor turned out to be the most accurate model when attempting to forecast future prices.

Article History
Article Received*: 25 March 2022*
Revised*: 30 April 2022*
Accepted*: 15 June 2022*
Publication*: 19 August 2022*

Keywords: Stock price, social media, machine learning, LSTM, decision tree, Random Forest, XG Boost regressor, and RMSE are some of the keywords that can be found in this article.

1 Introduction

The phrase "stock market" refers to the public marketplaces that are open for the issuance, acquisition, and selling of stock exchange. These marketplaces are also known as "bourses." There is a possibility that the financial market might be influenced by anything from a natural catastrophe or political event to a fundamental business decision

or even an event. The forecasting of the stock market has bazeen the topic of a significant amount of study, and in the process, a variety of models connected to time series analysis have been used.

The production of emotion characteristics may be helped along by methods used in sentiment machine learning. These algorithms take data from social networks and the news in order to do so. Aspects of the technology are also necessary, such as historical OHLC data spanning a variety of years that have been established and are used with the right technology. In recent years, machine learning (ML) strategies have gained popularity for use in time series research and stock market forecasting. This is largely attributable to the fact that these strategies are able to accurately predict future circumstances using massive historical datasets.

The most important addition made by this study is the use of online social networks rather than traditional forms of media in order to compile information about stock prices and corporate news. It is required to assess the tenor of each piece of news, and the data from Kaggle is utilized in order to develop emotion traits and merge them with technological aspects. The recommended machine learning model uses the emotional and technical aspects as inputs, and then it sends those attributes through a number of different models to determine which one has the greatest performance.

2 RELATED WORK

It is reasonable to conceive of today's stock markets as social networks since they have many members who contribute to their financial well-being. As a result, this comparison is appropriate. Players are faced with a difficult choice while deciding whether or not to sell or acquire shares. Investors are required to seek help from licensed financial planners, yet even these professionals struggle to make the best possible judgments for their clients' investments. When making choices, it may be quite advantageous to take into consideration the ideas of the many seasoned financial experts who speak out on the current status of the market and stock prices. Therefore, social media networks such as Twitter and Facebook, along with other news channels, are good places to look for financial information. This will improve the likelihood that information will be gathered and shared. Combining this information with real financial facts allows for more educated decisions to be made about investments.

For the purpose of predicting the stock market, a number of machine learning approaches, such as linear regression, logistic regression, and support vector machines (SVM), as well as certain deep learning models, such as RNN and LSTM, were used. Because more data were supplied all the way through the training phase of the model, it has earned a reputation for generating higher levels of accuracy with lower levels of loss.

3 MODEL OVERVIEW

For the purposes of prediction and classification of the stock market, sentiment analysis, together with four statistical machine learning models, such as Random Forest, LSTM, XG Boost Regressor, and Gradient Boost Classifier, are used. The two primary input

characteristics for the prediction model that is shown in Fig. 1 are the sentiment and the technical elements, which are combined during the training phase. These models are used to forecast stock prices or stock movements using the assistance of tweets gathered from Twitter and a company's stock market movements. This information is employed to create the forecast. Over the course of one day, we subjected the data from Twitter to a sentiment analysis to determine the impact that it had on stock market forecasts. RMSE numbers, which stand for root mean squared error, are used to evaluate and compare different models.

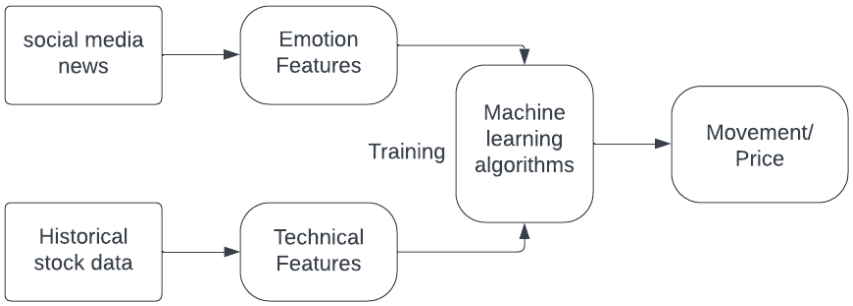

Fig. 1. Block Diagram of stock trend prediction

1 METHODOLOGY

1.1 Data collection

Data from [Kaggle] was used for the purpose of doing sentiment analysis on Apple stock. The beginning of the analysis will be "January 1, 2016," and it will end on "August 31, 2019." We utilized Yahoo Finance to compile daily data for the same time period for Apple shares so that we could evaluate historical data and make predictions about stock prices and trends. The data from Twitter was filtered such that it only included information pertaining to days on which the stock market was open (about 252 trading days per year). A final.csv file was generated using data from Twitter as well as daily stock prices, and this file served as the foundation for the analysis. For each and every one of our anticipated assessments, we made use of the modified closing price of AAPL.

1.2 Engineering of the features

The technique of prediction makes use of both technical and emotional components, each of which is broken down and described in further depth in the following subsections. For the purpose of gathering technical characteristics, the historical technical datasets of the selected companies in the stock market are employed. The formulations and explanations of the chosen technical characteristics are shown in Table 1. The dataset is analyzed to determine which qualities are the most significant**.**

Table 1.

Sr. No.	*Feature*	*Descriptions*
1	Opening Price (O)	The price of a stock at the opening of a trade day.
2	Volume	The daily volume of shares ex-changed.
3	Closing Price (C)	The price of a stock at the closing of a trade day.
4	Lowest Price (L)	Lowest price of stock during a day.
5	Highest Price (H)	Highest price of stock during a day.
6	Percentage change (Pct_change)	Percentage change in adjacent close value.

The technical characteristics are pre-processed in order to get them ready for usage in training. Methods such as feature scaling, normalization, and reshaping are used in the process of constructing a dataset. The Twitter dataset that was collected provided results in the form of emotion traits. Table 2 provides a listing of the psychological characteristics chosen for use in training the model.

Table 2.

Sr.No.	*Features*	*Descriptions*
1	Polarity(ts_polarity)	Polarity of twits
2	Volume(twitter_ val- ume)	Volume of the twits

Several machine learning methods were used to train and assess models in order to investigate and identify the stock trend prediction or movements of employing stocks twitter sentiment and stock price.

1.1 ALGORITHMS

In order to do the analysis and categorization, the following algorithms were utilized:

Classifier based on the Random Forest. The output of the random forest method, which is also a supervised learning technique that uses ensemble learning, is more accurate than that of a single classifier since it relies on a collection of classifiers rather than just one for its final output value. It may be used by classifiers as well as regressors. Bagging and boosting are the two sub-components that make up ensemble learning, and random forest is considered to be part of the bagging sub-component. When we use

bagging, we divide the dataset into multiple different training datasets, and give each classifier its own dataset to work with. Even while we can provide each classifier with just a single training dataset, it is possible that employing a variety of datasets can provide more accurate results. After reviewing the output of each classifier once again, it decides what the final result will be.

Working - It uses the decision tree approach, which includes selecting properties, features, or samples at random. This is done in order to maximize accuracy. If we already have a dataset, it will be converted into a Bootstrap dataset before it is used (It will create a new dataset using samples chosen at random). The property of the target will come next. The Random Forest algorithm makes use of a number of different Decision Trees, each of which is created in a completely arbitrary way. The characteristic that will be used for the root node is now chosen at random, and the root node will be chosen based on whatever attribute partitions the data the most effectively into subsets. Following this pattern, further root nodes are chosen, and so on, until the decision tree has been constructed. In addition, bootstrap data is used during the process of developing each and every decision tree.

Boosting Gradient Classifier. The Gradient Booster method builds trees one at a time, with each newly constructed tree contributing to the process of correcting the mistakes made by previously trained trees. When more trees are added to the model, the level of expression increases. Each every tree that is built is shallow in general, and there are three factors to take into account: learning pace, number of trees, and depth of trees.

Estimators of the learning rate, as well as the number of them: The efficiency and precision of a model are both affected by a learning algorithm's hyperparameters, which are crucial components of the algorithm. The learning rate and the n estimators are two hyperparameters that are absolutely necessary for gradient boosting decision trees. When we talk about a model's "learning rate," we're referring to the pace at which it can absorb new information. When new trees are added, the general structure of the model is altered. The magnitude of the change is directly proportional to the pace of learning. When the learning rate is decreased, the model will learn at a more gradual pace. Because of the slower learning rate, the model ends up being more dependable and efficient overall. When it comes to statistical learning, models that learn more slowly do better. On the other side, a high cost is associated with slow learning. This brings us to the discussion of the second significant hyperparameter, which requires a longer period of time to train the model.

LSTM RNN. Long-Short-Term Memory networks, or LSTM networks for short, are a type of RNN that have become increasingly popular for use in time series and sequential analysis. Every one of them is made up of three layers: a hidden layer, an output layer, and an input layer. Memory cells and three gates that are in charge of updating the cell state make up the LSTM network's hidden layer. This layer is part of an LSTM network. The vanishing gradient issue, which affects RNNs but not LSTM networks, is crucial for the prediction of stock prices because the processing of previous data by the neural network influences subsequent and future data. LSTM networks are immune to this problem. Figure 2 illustrates the LSTM network's general topology in addition to its

three gates.

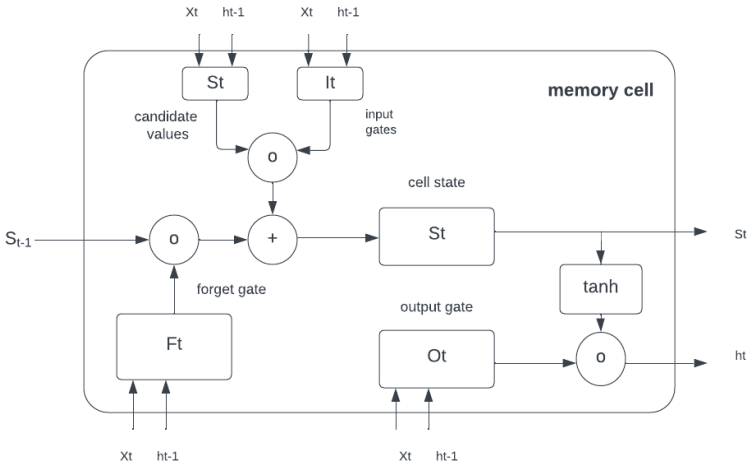

Fig. 2. Long Short Term Memory Network

The values of the updated cell state are determined by the gates' output. The following equations depict this,

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (2)$$

$$c_t = tanh(W_c . [h_{t-1}, x_t] + b_c) \quad (3)$$

$$o_t = sigma(W_o . [h_{t-1}, x_t] + b_o) \quad (4)$$

$$h_t = sigma * tanh(c_t) \quad (5)$$

where x_t, h_t and f_t are input, output vectors and vector depicting the forget gate; c_t is a representation of the cell state vector; it is a vector of the input gate; o_t is a vector of the output gate; and W, b are matrices and vectors representing the parameter values.

XG Boost Regressor. Extreme Gradient Boosting (XG Boost) is a distributed, scalable gradient-boosted decision tree (GBDT) machine learning framework. In supervised machine learning, a model is trained using algorithms to discover patterns in a dataset of features and labels, and the model is then used to predict the labels on the features of a new dataset. It attempts to optimize a cost objective function made up of a regularization term (β) and a loss function (d)

$$\Omega(\theta) = \sum^{n} d(y \quad \hat{y}) + \sum^{K} \beta(f) \quad (6)$$
$$\underbrace{i = 1}\ i \rightarrow \quad \underbrace{k = 1}\ k \rightarrow$$
$$loss\ function \qquad regularisation$$

where K, f_k, n and $\hat{y}$ is a tree from the ensemble trees, number of trees to be generated, the number of instances in the training set and the predictive value respectively. The following is a definition of the regularization term:

$$(f) = \gamma T + \frac{1}{} [\alpha \sum^{T} |c| + \lambda \sum^{T} c^{2}] \quad (7)$$

$$t \qquad {}_2 j=1 \quad j \; j=1$$

where $c \; and \; \lambda$ are the weight associated with each leaf and regularization term on the weight, and γ is the minimum split loss reduction. Let's say that $f_t(x_i) = C(x_i)$ with q in the range [1, T], where T is the number of leaves.

By evaluating the error between the predicted values and observed values, the accuracy of the model may be verified. It is recommended to use root mean square error, mean square error, and mean absolute error. The model performs better and more accurately as the output is closer to zero. MAE, MSE, and $RMSE$ are each calculated using following equations.

$$MAE = {}^{1}\sum|y - \hat{y}| \qquad (8)$$

$$MSE = {}^{1}\sum_{n}^{2}(y - \hat{y})^2 \qquad (9)$$

$$n \quad i=1$$

$$\overline{RMSE} = \sqrt{MSE} \qquad (10)$$

1 RESULTS

Results for classification using gradient boosting classifier is shown in Fig. 3.: It gives accuracy of 55%

Classification Report

	precision	recall	f1-score	support
0	0.52	0.21	0.30	127
1	0.56	0.83	0.67	150
accuracy			0.55	277
macro avg	0.54	0.52	0.48	277
weighted avg	0.54	0.55	0.50	277

Fig. 3. Gradient boosting classifier classification report

Result for classification using Random Forest classifier is shown in Fig. 4.: It gives 53% accuracy.

Classification Report

	precision	recall	f1-score	support
0	0.48	0.29	0.36	127
1	0.55	0.73	0.63	150
accuracy			0.53	277
macro avg	0.52	0.51	0.50	277
weighted avg	0.52	0.53	0.51	277

Fig. 4. Random Forest classifier classification report

The classification approach could not produce enough confidence to justify use to correlate stock movement with Twitter sentiments, with accuracy scores in the range of 52–56 percent.

Results of regression using Random Forest regressor shown in Fig. 5.: The following figure shows the graph between actual data and predicted data. It gives RMSE value equals to 0.0993 and R-squared value equals to 82.27%.

Results of regression using LSTM is shown in Fig. 6.: The following figure shows the graph between actual data and predicted data. It gives RMSE value equals to 0.1334 and R-squared value equals to 68.01%.

Results of regression using XG Boost is shown in Fig. 7.: The following figure shows the graph between actual data and predicted data. It gives RMSE value equals to 0.0477 and R-squared value equals to 95.92%.

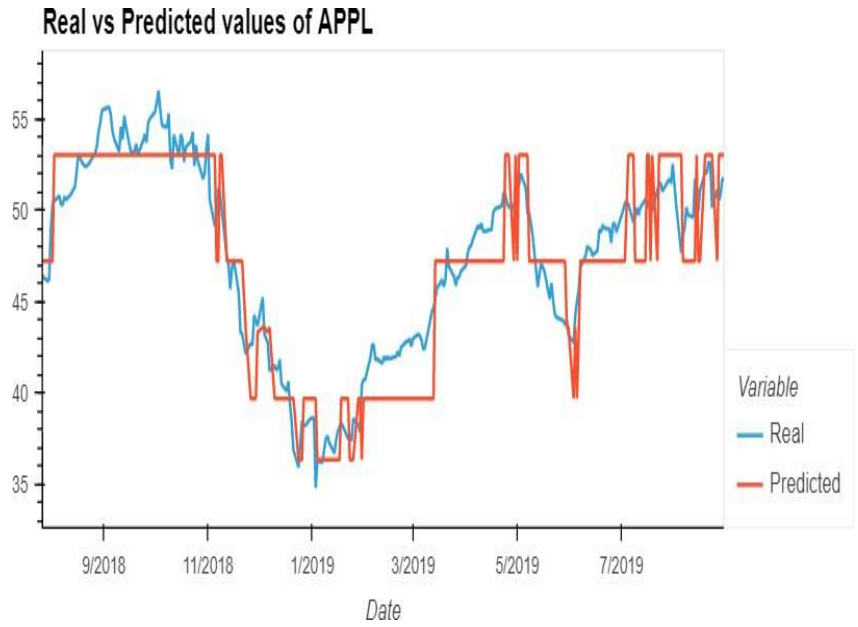

Fig. 5. Real and predicted value of APPL using Random Forest Regressor

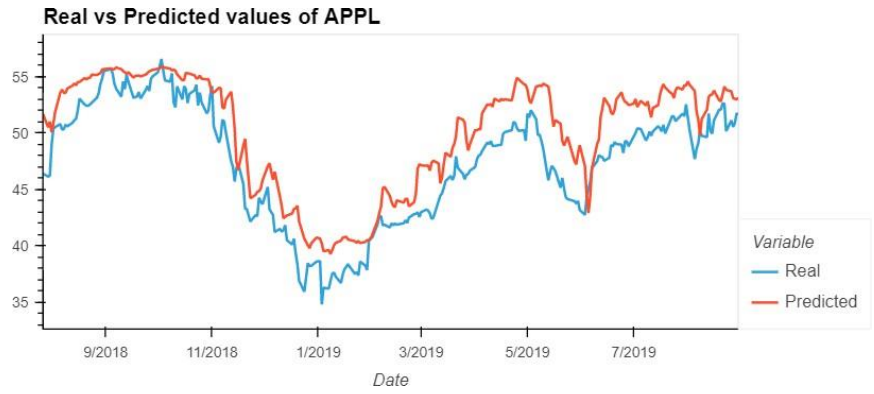

Fig. 6. Real and predicted value of APPL using LSTM

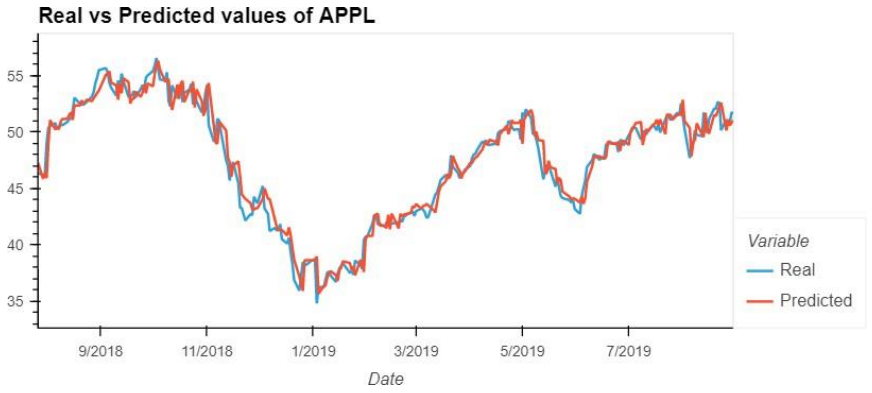

Fig. 7. Real and predicted value of APPL using XG Boost

CONCLUSION

In this particular investigation, we make use of a variety of approaches to price prediction. These include random forest, LSTM, and XG Boost, in that particular sequence. Using the root mean squared error (RMSE), we determined the magnitude of the discrepancy between the anticipated price and the actual price. In comparison to LSTM and random forest, XG Boost delivered the most accurate results. This is because the model will be more accurate the closer a value was to zero. The price in real time was anticipated really well by the graph that was generated by XG Boost.

Because its error was limited, XG Boost outperforms random forest and LSTM in terms of efficiency, accuracy, and outcomes. This is because its error was limited, which makes it very ideal for predicting the real stock price and reducing the uncertainty of future value. As a result, it can be said that XG Boost outperforms random forest and LSTM in terms of efficiency, accuracy, and outcomes.

References

1. Diren Archary, Marijke Coetzee: Predicting Stock Price Movement with social media and Deep Learning. 2020 International Conference on Artificial Intelligence, Big Data, Compu- ting and Data Communication Systems (icABCD), 2020
2. Samuel Olusegun Ojo, Pius Adewale Owolawi, Maredi Mphahlele, Juliana Adeola Adisa.: Stock Market Behaviour Prediction using Stacked LSTM Networks. 2019 International Multidisciplinary Information Technology and Engineering Conference (LIMITED), 2019.
3. Iyad Lahsen Cherif, Abdesselem Kortebi.: On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification. 2019 Wireless Days (WD), 2019
4. Vazirani, Sahil, Abhishek Sharma, and Pavika Sharma.: Analysis of various machine learn- ing algorithm and hybrid model for stock market prediction using python. In 2020 Interna- tional Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pp. 203-207. IEEE, 2020.
5. Ranibaran, Golshid, Mohammad-Shahram Moin, Sasan H. Alizadeh, and Abbas Koochari.: Analyzing effect of news polarity on stock market prediction: a machine

learning approach In 2021 12th International Conference on Information and Knowledge Technology (IKT), pp. 102-106. IEEE, 2021.
6. Lakshminarayanan, Sai Krishna, and John P. McCrae.: A Comparative Study of SVM and LSTM Deep Learning Algorithms for Stock Market Prediction. In AICS, pp. 446-457. 2019.
7. Rohatgi, Sachin, Krishna Kumar Singh, and Deepmala Jasuja.: Comparative Analysis of Machine Learning Algorithm to Forecast Indian Stock Market. In 2021 International Con- ference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 278-283. IEEE, 2021.
8. Sun, Linyu.: Application and improvement of Xgboost algorithm based on multiple param- eter optimization strategy. In 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), pp. 1822-1825. IEEE, 2020.
9. Ma, Zhong, Jiansheng Guo, Sheng Mao, and Taoyong Gu.: An Interpretability research of the Xgboost algorithm in remaining useful life prediction. In 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), pp. 433-438. IEEE, 2020.
10. Nabipour, Mojtaba, Pooyan Nayyeri, Hamed Jabani, S. Shahab, and Amir Mosavi.: Predict- ing stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. IEEE Access 8 (2020): 150199-150212.
11. Lin, Yaohu, Shancun Liu, Haijun Yang, and Harris Wu.: Stock Trend Prediction Using Can- dlestick Charting and Ensemble Machine Learning Techniques With a Novelty Feature En- gineering Scheme. IEEE Access 9 (2021): 101433-101446.
12. Li, Qing, Jinghua Tan, Jun Wang, and Hsinchun Chen.: A multimodal event-driven lstm model for stock prediction using online news. IEEE Transactions on Knowledge and Data Engineering 33, no. 10 (2020): 3323-3337.
13. Kilimci, Zeynep Hilal, and Ramazan Duvar.: An efficient word embedding and deep learn- ing based model to forecast the direction of stock exchange market using Twitter and finan- cial news sites: a case of Istanbul stock exchange (BIST 100). IEEE Access 8 (2020): 188186-188198.
14. Alsubaie, Yazeed, Khalil El Hindi, and Hussain Alsalman.: Cost-sensitive prediction of stock price direction: Selection of technical indicators. IEEE Access 7 (2019): 146876-146892.