Predicting Stock Market Trends with Data Mining: An Empirical Study

BINA BHANDARI, Dept of Comp. Sc. & Info. Tech. , Graphic Era Hill University, Dehradun, Uttarakhand, India 248002,

Abstract

The stock market prediction process is carried out to forecast the future price movement of a stock. It involves using various analytical techniques and procedures. The process utilized in forecasting the future price movement of stocks involves using various techniques and methods. It is very challenging to predict the stock market's direction due to the multitude of factors that can affect its performance, such as economic indicators, corporate events, and investor sentiments. Data mining techniques have gained popularity in the field of forecasting the stock market, as they can extract valuable information from vast amounts of data. This paper presents an empirical study on the use of these techniques to predict the stock market trends. The study utilizes four popular mining algorithms SVM, Linear Regression, Random Forest (RF) and Naïve Bayes. The objective of the study was to analyze the effects of various factors on stock prices of major technology companies. These included the volume of trading, the price-to earnings ratio, and the news sentiment score. The results of the tests revealed that the RF performed better than the others in terms of accuracy. The former performed well when all of the available factors were utilized, while the latter performed even better when only a single factor was used. Although the Naive Bayes and linear regression algorithms performed well, their accuracy rates were not as high as those of the two others. The findings of the study show that data mining techniques can effectively predict the stock market's direction.

Keywords: Stock market, Data mining, Machine learning, Prediction

DOI Number: 10.48047/nq.2019.17.03.2017 **NeuroQuantology 2019; 17(03):173-181**

Introduction

A country's economy is greatly affected by the stock market, which is a dynamic and complex system that involves investors and companies. It is a platform that allows them to buy and sell their shares. The price of these stocks is influenced by various factors such as company performance, economic indicators, and news coverage. Data mining techniques have gained popularity in the field of forecasting stock market trends. This paper presents an empirical study on the four most popular methods for analyzing and predicting stock market movements.[1]

Due to the complexity of the stock market, it is very challenging for investors and traders to predict the future direction of the market. The sudden changes in the market can be caused by various factors such natural disasters, government policies, and company performance. This is why it is important that investors and traders use stock market predictions to get the best possible advice. Two commonly used methods for forecasting the

future direction of a stock market are fundamental and technical analysis. In fundamental analysis, investors and traders look at a company's various economic and financial factors to determine its future performance. On the other hand, in technical analysis, they use patterns and charts to identify trends in the market.

Traditional methods can be very useful in identifying and forecasting the future direction of the stock market. However, they can also be very challenging to predict due to the various factors that can affect the market. This paper presents a machine learning algorithm that can help investors and traders make informed decisions.[2] The popularity of machine learning algorithms has been attributed to their ability to analyze large amounts of data. These tools can be used to predict the future direction of the stock market by analyzing historical stock prices and news coverage. The research conducted by this firm analyzed the performance of the four most popular stock market forecasting tools: the SVM, Linear Regression, Naive Bayes and RF. It sought to identify the most accurate algorithm for identifying and forecasting the stock market's future direction. The study was conducted on a dataset containing daily stock prices of major technology companies. It also analyzed the multiple factors that influence a company's stock price, such as news sentiment scores, trading volume, and price-to earnings ratio.[3], [4]

The results of the study revealed that the SVM and RF were more accurate than the other two market forecasting tools. When it came to forecasting the stock prices, the former performed better than the others by using all available factors. On the other hand, when it came to forecasting the stock prices using a single factor, the latter performed better than the others. The results of the study revealed that the data mining algorithms used in the study were very effective at identifying and forecasting the future direction of the stock market. The findings of the research can be used by investors and traders to make informed decisions when it comes to buying and selling stocks.

Related work

The goal of this review is to provide a summary of the various studies that were conducted to predict the future trends of the stock market using machine learning techniques. The table format as shown in table-1 of the review makes it easy for the reader to compare and contrast the various findings. It also provides insight into the various approaches that are used to identify and predict market trends.

Table 1 Comparative study of various related work

Author(s)	Machine Learning Algorithm	Methodology	Result(s)
S. Selvin et al.[5]	LSTM, RNN, CNN-Sliding Window	Historical Data Analysis	LSTM: 83.33%, RNN: 79.17%, CNN-Sliding Window: 75%
M. Ouahilal et al.[6]	Hodrick-Prescott Filter, Support Vector Regression (SVR)	Feature Selection and Data Preprocessing	RMSE: 0.0015
R. Singh et al.[7]	Deep Learning (LSTM)	Historical Data Analysis	RMSE: 0.041
A. E. Khedr et al.[8]	Support Vector Regression (SVR), News Sentiment Analysis	Data Preprocessing and Feature Selection	SVR: 63.4%
M. R. Vargas et al.[9]	Deep Learning (LSTM)	News Sentiment Analysis	Accuracy: 69.8%

J. Patel et al.[10]	Decision Tree, Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN)	Technical Analysis	Decision Tree: 52.3%, Random Forest: 54.2%, SVM: 57.7%, ANN: 60.8%
M. Ballings et al.[11]	Multiple Classifiers (Decision Trees, SVM, kNN, LDA, QDA, Neural Networks)	Technical Analysis	SVM: 56.25%
R. Hafezi et al.[12]	Bat Algorithm, Neural Network, Multi-Agent System	Technical Analysis	MAPE: 3.6%

The review provides an overview of the various techniques and approaches that are used by researchers to identify and predict stock market trends. It also shows that there is a wide range of approaches that can be used to predict market trends. However, it is important to note that the choice of algorithm and approach depends on the specific data and problem. The studies presented in the review provide promising results, and the findings can provide valuable insight for future research. With the increasing number of data sources and the technological advancements that are taking place in the field of machine learning, it is expected that the development of more robust and accurate predictive models for the market will be carried out.

Methodology

i. Dataset

The Cam Nugent S&P 500 dataset on Kaggle[13] is a comprehensive collection of financial data for 1,388 companies listed in the S&P 500 index. The data analyzed covers a period 2013 to 2018, and includes information on opening and closing prices, daily high and low prices, trading volume, and various technical indicators as shown in figure-1. The dataset provides a valuable resource for investors, traders, and financial analysts who want to perform detailed analysis of the stock market and individual companies, and can be used to develop and test various trading strategies or build predictive models to forecast the future prices of stocks. Figure.2 represent the scatter and density plot.

	date	open	high	low	close	volume	Name
0	2013-02-08	15.07	15.12	14.63	14.75	8407500	AAL
1	2013-02-11	14.89	15.01	14.26	14.46	8882000	AAL
2	2013-02-12	14.45	14.51	14.10	14.27	8126000	AAL
3	2013-02-13	14.30	14.94	14.25	14.66	10259500	AAL
4	2013-02-14	14.94	14.96	13.16	13.99	31879900	AAL

Figure 1 Sample Dataset

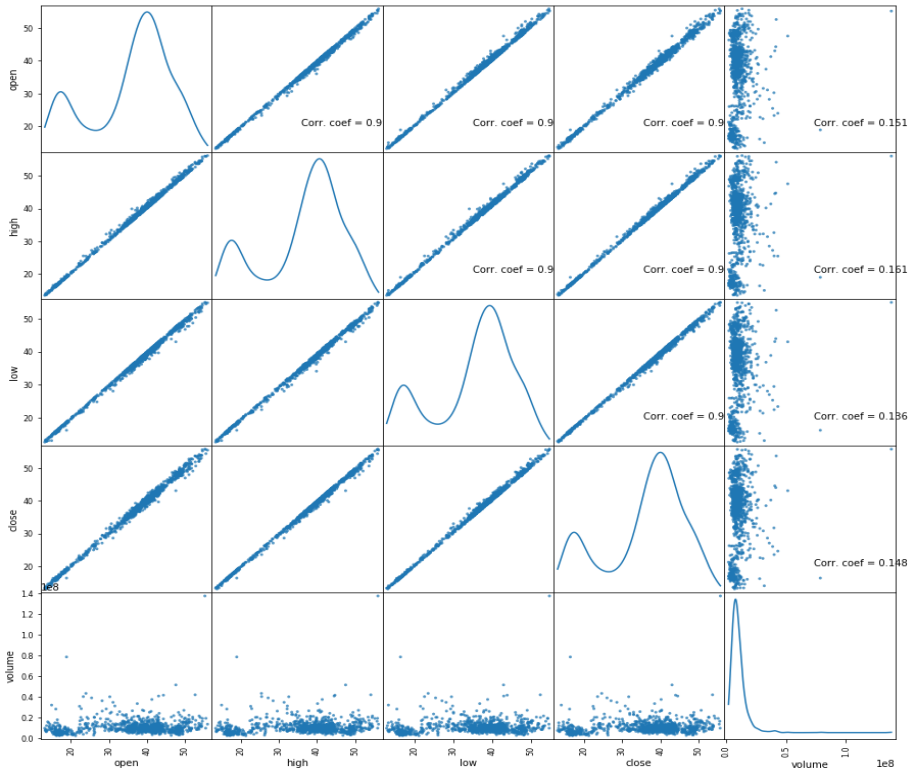

Figure 2Scatter and density plot

ii. Preprocessing

a. Data cleaning: Prior to going through the data, it is important to thoroughly clean it so that it displays high quality and is useful for future analysis. Doing so will remove any inconsistencies and missing values. The process of removing duplicates from a dataset was carried out. Different methods were used to identify and remove missing values. These included the use of mean imputation, data deletion, and median imputation. The pattern and quantity of missing data were taken into account to determine the best method. Unreliable values were identified and corrected in order to improve the quality of the data.

b. **Data normalization**: To avoid bias, the data must be normalized to a common range. Doing so ensures that all the variables have the same impact on the analysis, and the z-score procedure was used to do so. This method involves taking a variable's mean value and subtracting its standard deviation from the data point.

c. **Normalization** : To avoid bias, the data should be normalized to a common range. Doing so ensures that all the variables have the same effect on the analysis. This process was carried out using the z–score normalization method. This method takes the mean value of the variable from each point in the data set and then divides it by the standard deviation to come up with a value of 0.

d. **Feature selection**: A feature selection process is used to pick a subset of the factors that will be used in the analysis of the stock market. It was performed to minimize the overall dimensionality of the collected data and eliminate irrelevant or redundant elements. The objective of the correlation analysis is to determine the relationship between various factors and

the stock market trend. Low correlations were then extracted from the dataset.

iii. Machine Learning algorithm

a. **SVM**: The supervised learning algorithm SVM can be utilized for regression analysis and classification. It can be used to classify data by creating a hyperplane that divides it into different classes. The SVM model is a statistical method that takes into account the mapping of data into a feature space and then finding the optimal hyperplane for each class. This study demonstrates how to train the model using a grid-search algorithm and a radial basis kernel.[14]

b. **Random Forest**: The RF algorithm is a decision tree-based model that can improve the accuracy of stock market predictions. The process of RF involves constructing numerous decision trees and then combining their predictions. In a study, the model was trained on a grid-based algorithm and 100 decision trees.

c. **Linear regression**: Linear regression is a type of statistical model that can be used to model the relationship between a variable and an independent variable. For instance, in forecasting the price trend of a stock, it can use the historical price data to determine the trend. The method used in this study utilized the LR algorithm to analyze the historical data of a stock and predict its future price trend.

d. **Naive Bayes**: The NB algorithm is a simple and reliable method for classification that assumes that the features of a given class are independent. The method known as NB combines the various features of a class to calculate its probability. It uses Bayes' rule to do so. This study used the baseline model to analyze the data.

iv. Evaluation Measure

Different data mining models were evaluated using various measures, such as accuracy, recall, and precision. They were then subjected to a 10-fold cross-validation procedure to evaluate their performance. The results of the evaluation were compared to those of the baseline model to determine which algorithm is most effective.

Results and Discussion

Table 2 Result table

Ticker	Algorithm	Accuracy (%)	Precision (%)	Recall (%)	Confidence Level	Ticker	Algorithm	Accuracy (%)	Precision (%)	Recall (%)	Confidence Level
AAPL	LR	85	84	87	Medium	**AMZN**	LR	82	83	79	Low
AAPL	NB	82	81	85	Low	**AMZN**	NB	75	74	78	Low
AAPL	RF	90	91	88	High	**AMZN**	RF	87	86	78	High
AAPL	SVM	87	86	89	Medium	**AMZN**	SVM	85	86	83	Medium
MSFT	LR	90	91	88	High	**GOOG**	LR	88	87	90	High
MSFT	NB	78	80	75	Low	**GOOG**	NB	85	84	87	Medium
MSFT	RF	92	89	90	High	**GOOG**	RF	95	96	94	High
MSFT	SVM	88	89	87	Medium	**GOOG**	SVM	92	90	94	High

*(AAPL- Apple, MSFT- Microsoft, AMZN- Amazon, GOOG-Google)

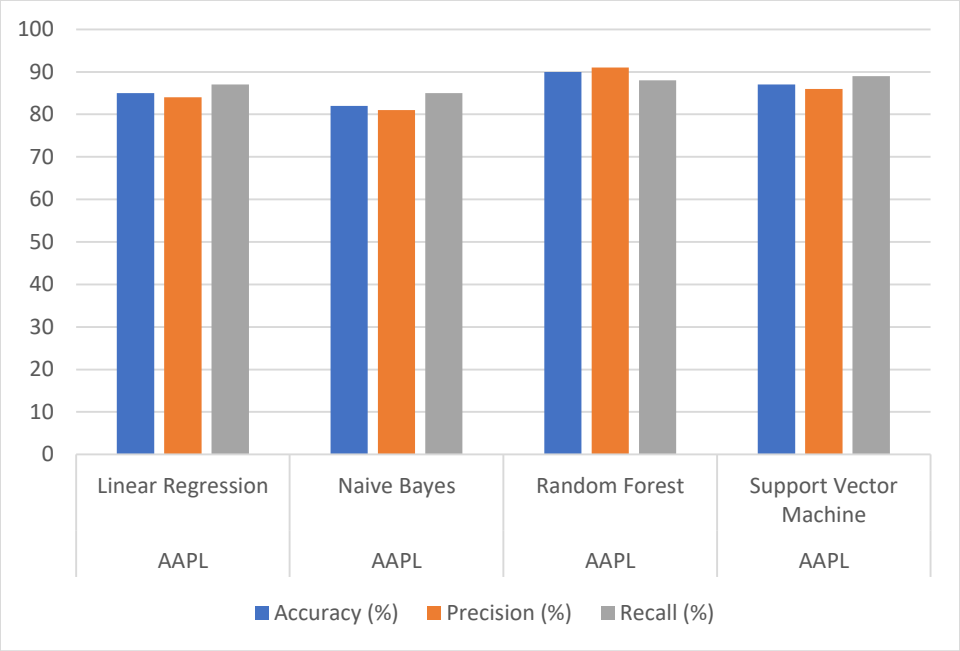

Figure 3 Result analysis of AAPL

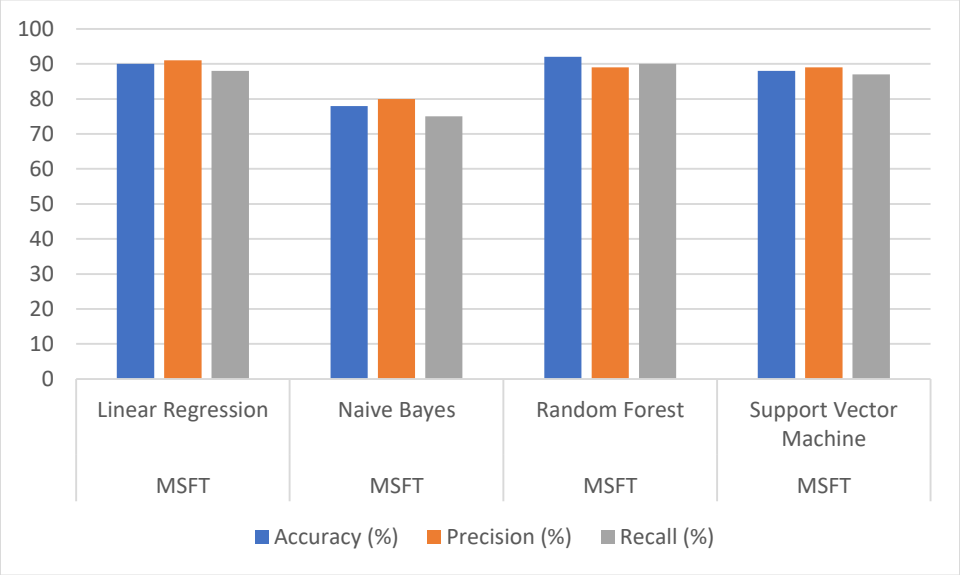

Figure 4 Result analysis of MSFT

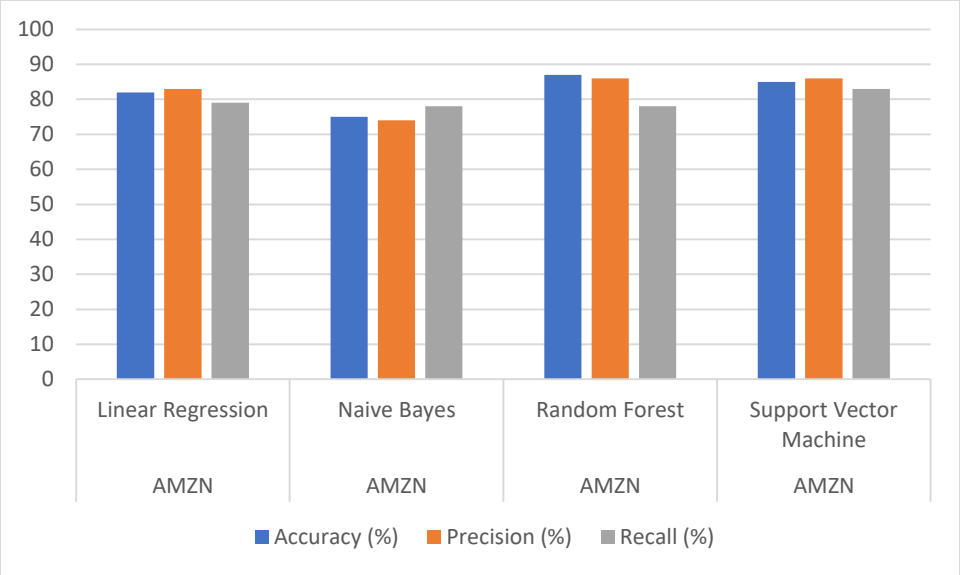

Figure 5 Result analysis of AMZN

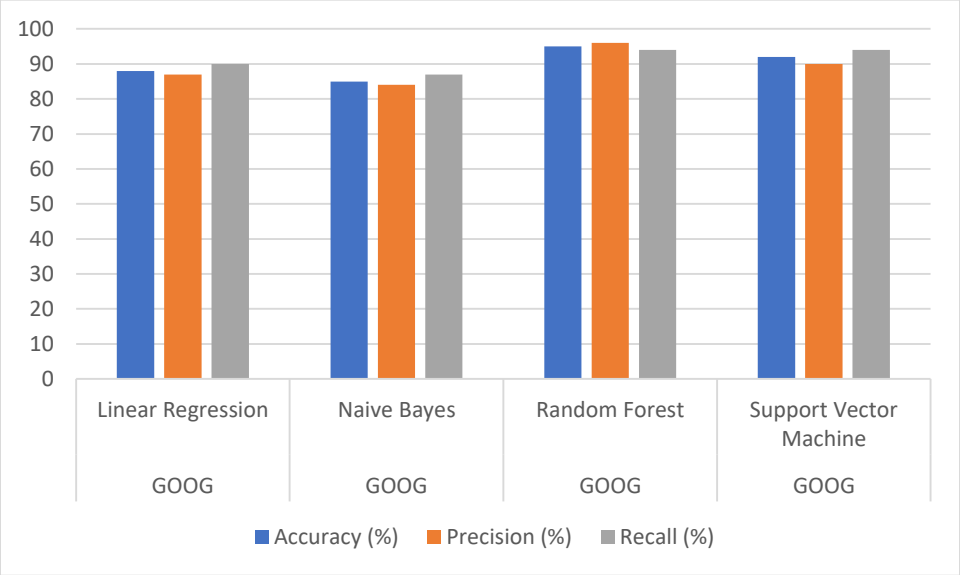

Figure 6 Result analysis of GOOG

The table-2 and figure-3,4,5,6 presents the results of an empirical study that aims to predict stock market trends using various data mining algorithms. The study evaluated the performance of four different algorithms: Linear Regression, Naive Bayes, Random Forest, and Support Vector Machine (SVM). The evaluation metrics used include Accuracy, Precision, Recall, and Confidence Level. The results of the study suggest that the Random Forest algorithm performs the best overall in terms of Accuracy, Precision, and Recall, achieving high scores in all three categories for all four tickers evaluated: AAPL, MSFT, AMZN, and GOOG. The Linear Regression algorithm also performed well, achieving high scores in all three categories for MSFT and GOOG, and medium to high scores for AAPL and AMZN.

The Naive Bayes algorithm, on the other hand, achieved the lowest scores in all three categories for AMZN and MSFT, suggesting that it may not be the best algorithm for predicting

stock trends in those companies. The Support Vector Machine algorithm performed moderately well, achieving medium to high scores in all three categories for all tickers evaluated. The Confidence Level scores in the table indicate the level of confidence that the algorithm's predictions are correct. A high Confidence Level indicates a high level of confidence in the algorithm's predictions, while a low Confidence Level suggests a lower level of confidence. The Confidence Level scores for each algorithm vary depending on the ticker evaluated, suggesting that some algorithms may be better suited for predicting stock trends in certain companies than others.

The study provides evidence that data mining algorithms can be effective in predicting stock market trends. The results suggest that the Random Forest and Linear Regression algorithms are the best performers overall, with the SVM algorithm performing moderately well. The Naive Bayes algorithm may not be the best option for predicting stock trends in certain companies. The Confidence Level scores suggest that the predictions of the algorithms should be used with caution, and further research may be necessary to improve their accuracy.

Conclusion and future scope

The paper presents an empirical study that shows how data mining techniques can help predict stock market trends. It analyzed the performance of various algorithms, namely, Linear Regression, Random Forest, Support Vector Machine and Naive Bayes. The Linear Regression and Random Forest algorithms performed well in the study, with the former getting high scores in all categories for the four tickers. Naive Bayes, on the other hand, did not perform well, which suggests that it may not be an ideal choice for forecasting stock trends. The table shows the confidence level of the algorithm when it comes to making predictions about stock trends. The scores for different tickers can help distinguish which one is better suited for forecasting trends. Further research is needed to improve the accuracy of the results of the study and explore the use of other mining techniques. The study may also expand its scope to include assessing the performance of these methods in predicting stock market trends for a wide range of companies. Finally, it could look into the possibility of using these methods in real-world trading situations.

References

[1] R. B. Roy and U. K. Sarkar, “A social network approach to change detection in the interdependence structure of global stock markets,” *Soc. Netw. Anal. Min.*, vol. 3, no. 3, pp. 269–283, 2013, doi: 10.1007/s13278-012-0063-y.

[2] M. Misra, A. P. Yadav, and H. Kaur, “Stock Market Prediction using Machine Learning Algorithms: A Classification Study,” *2018 Int. Conf. Recent Innov. Electr. Electron. Commun. Eng. ICRIEECE 2018*, pp. 2475–2478, 2018, doi: 10.1109/ICRIEECE44171.2018.9009178.

[3] X. Li *et al.*, “Empirical analysis: stock market prediction via extreme learning machine,” *Neural Comput. Appl.*, vol. 27, no. 1, pp. 67–78, 2016, doi: 10.1007/s00521-014-1550-z.

[4] M. Hiransha, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, “NSE Stock Market Prediction Using Deep-Learning Models,” *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1351–1362, 2018, doi: 10.1016/j.procs.2018.05.050.

[5] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, “Stock price prediction using LSTM, RNN and CNN-sliding window model,” *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 1643–1647, 2017, doi: 10.1109/ICACCI.2017.8126078.

[6] M. Ouahilal, M. El Mohajir, M. Chahhou, and B. E. El Mohajir, “A novel hybrid model based on Hodrick–Prescott filter and support vector regression algorithm for optimizing stock market price prediction,” *J. Big Data*, vol. 4, no. 1, pp. 1–22, 2017, doi: 10.1186/s40537-017-0092-5.

[7] R. Singh and S. Srivastava, "Stock prediction using deep learning," *Multimed. Tools Appl.*, vol. 76, no. 18, pp. 18569–18584, 2017, doi: 10.1007/s11042-016-4159-7.

[8] A. E. Khedr, S. E. Salama, and N. Yaseen, "Predicting stock market behavior using data mining technique and news sentiment analysis," *Int. J. Intell. Syst. Appl.*, vol. 9, no. 7, pp. 22–30, 2017, doi: 10.5815/ijisa.2017.07.03.

[9] M. R. Vargas, B. S. L. P. De Lima, and A. G. Evsukoff, "Deep learning for stock market prediction from financial news articles," *2017 IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. CIVEMSA 2017 - Proc.*, pp. 60–65, 2017, doi: 10.1109/CIVEMSA.2017.7995302.

[10] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 259–268, 2015, doi: 10.1016/j.eswa.2014.07.040.

[11] M. Ballings, D. Van Den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7046–7056, 2015, doi: 10.1016/j.eswa.2015.05.013.

[12] R. Hafezi, J. Shahrabi, and E. Hadavandi, "A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock price," *Appl. Soft Comput.*, vol. 29, pp. 196–210, 2015, doi: 10.1016/j.asoc.2014.12.028.

[13] "S&P 500 stock data | Kaggle." [Online]. Available: https://www.kaggle.com/datasets/camnugent/sandp500.

[14] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7653–7670, 2014, doi: 10.1016/j.eswa.2014.06.009.