ISSN Online: 2821-1936

Transactions on Data Analysis in Social Science

Journal Homepage: https://transoscience.ir

Feature Extraction and Classification Methods for Stock Market Trend Prediction

G. Ranjbaran[1], M.S. Moin[2,*]

[1] Department of Computer Engineering, Faculty of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2] Associate Professor, Information Technology Research Institute, Research Institute for ICT (ITRC), Tehran, Iran

ARTICLE INFO

Article History:
Received 4 February 2019
Received in revised form 12 March 2019
Accepted 28 September 2019
Available online 29 September 2019

Keywords:
Stock Market; Stocks; Artificial Intelligence; Fundamental Analysis; Technical Analysis; Behavioral Analysis

ABSTRACT

Today, the stock market has become a crucial channel for mobilizing investors' capital. As a key indicator of a nation's economic and financial activities, the stock exchange plays a pivotal role in reflecting the overall economic performance of a country or region. Predicting stock price movements remains one of the most challenging tasks in the financial domain. Accurate stock prediction not only enhances investors' profitability but also contributes to national economic growth. Given the dynamic, complex, nonlinear, and nonparametric nature of stock markets, precise forecasting of stock price variations is essential for developing effective trading strategies. Researchers have employed various methodologies for stock market prediction, among which feature extraction and classification constitute the two fundamental processes. This study reviews and analyzes different feature extraction methods categorized into four types and classification techniques applied in prior research using artificial intelligence and mathematical models. The findings indicate that, due to the nonlinear nature of financial data, neural networks, particularly those employing hybrid or ensemble feature extraction approaches, demonstrate the highest efficiency and predictive performance in stock market forecasting.

1. INTRODUCTION

The stock exchange is a marketplace where shares can be traded or transferred. Today, the stock market has become a crucial channel for mobilizing investors' funds. On one hand, through the issuance of shares, a large amount of capital flows into the stock market, strengthening the organic composition of corporate capital, enhancing capital concentration, and promoting economic growth. On the other hand, the circulation of shares enables the effective aggregation of financial resources and fosters capital accumulation. Therefore, the stock market is regarded as a key indicator of economic and financial activity in a country or region. In particular, stock transaction prices

* Corresponding Author: moin@itrc.ac.ir
Associate Professor, Information Technology Research Institute, Research Institute for ICT (ITRC), Tehran, Iran

http://dx.doi.org/10.47176/TDASS.2019.163

often serve as indicators of stock value and quantity because they objectively reflect the market's supply and demand relationship [1].

Studying stock price prediction can guide investors toward profitable investments and, beyond individual gains, plays an essential role in national economic development. Stock price forecasting is an attractive field for both investors and researchers seeking to participate in financial markets [2]. However, predicting stock price movements remains one of the most challenging tasks in modern finance [3]. Given the dynamic, complex, nonlinear, and nonparametric nature of stock markets [4], accurate forecasting of stock price fluctuations is critical for developing effective trading strategies [5].

Researchers have proposed a variety of approaches for stock market prediction, leading to diverse results. In this study, a comprehensive review is conducted on feature extraction and classification methods employed for stock price prediction, categorized into linear and nonlinear approaches. Nonlinear methods are predominantly based on artificial intelligence techniques.

The remainder of this paper is structured as follows:

Section 2 discusses the general stages involved in stock market prediction.

Section 3 presents various methods and analytical techniques used for feature extraction in model development.

Section 4 introduces classification models applied in stock price forecasting.

Finally, Section 5 provides the conclusion and future research directions.

2. STAGES OF STOCK MARKET PREDICTION

Stock market prediction is essentially a data mining process that seeks to identify hidden patterns within large volumes of financial data to forecast market trends with acceptable accuracy. Similar to other data mining applications, stock forecasting involves several systematic stages, as illustrated in Figure 1.

It is worth noting that there is no standardized dataset for stock market prediction research. Different stock exchanges typically release their trading data publicly; however, even when two studies use data from the same stock market, the time intervals analyzed are often different. As a result, the datasets used across studies vary significantly in both scope and temporal coverage, leading to challenges in model comparison and reproducibility.

Fig. 1. Different Stages of Stock Market Prediction Processing

The block diagram presented in Figure 1 illustrates the stages that are generally followed in most classification problems. In this study, we focus on two main stages of this process feature extraction and classification as the core components of stock market prediction.

3. FEATURES USED IN THE STOCK MARKET

Due to its inherent appeal to both researchers and investors, the stock market has been extensively examined using a variety of methods to extract patterns and useful features aimed at forecasting market fluctuations. As discussed in Section 1, the stock market's high complexity and dynamic nature make it impossible to account for all influencing factors or extract every significant feature. However, as shown in Figure 2, the stock market problem can be analyzed from several different feature perspectives. These include technical, fundamental, behavioral, and hybrid features, which together constitute the necessary inputs for the classification process.

Before the emergence of soft computing techniques in stock prediction, studies primarily relied on two types of features technical and fundamental. Technical features are derived solely from historical price data, while fundamental features are based on macroeconomic indicators and the financial information of the issuing companies.

Technical features typically include information such as the highest and lowest prices at which a stock has been traded, the opening and closing prices, and the daily trading volume. Examples of such features include the Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), and Rate of Change (ROC) indicators [11].

Fundamental features, on the other hand, involve financial and statistical data published by companies on a quarterly, monthly, or annual basis. In this context, researchers seek to estimate the intrinsic value of company stocks. Additionally, macroeconomic factors such as oil prices, foreign exchange rates, and gold prices have a direct impact on stock market performance.

With advances in machine learning, behavioral features have recently been introduced alongside the traditional ones to improve prediction accuracy. In this category, forecasts are made based on user behaviors across social networks and media platforms such as Twitter and news agencies. It has been repeatedly observed that the dissemination of certain news or events can provoke emotional reactions among individuals, leading to impulsive buying or selling behaviors that disrupt market equilibrium by altering supply and demand dynamics. Therefore, features derived from behavioral analysis can be considered highly critical.

Lastly, hybrid features represent a combination of the aforementioned categories, integrating the advantages of previous methods to enhance overall prediction performance.

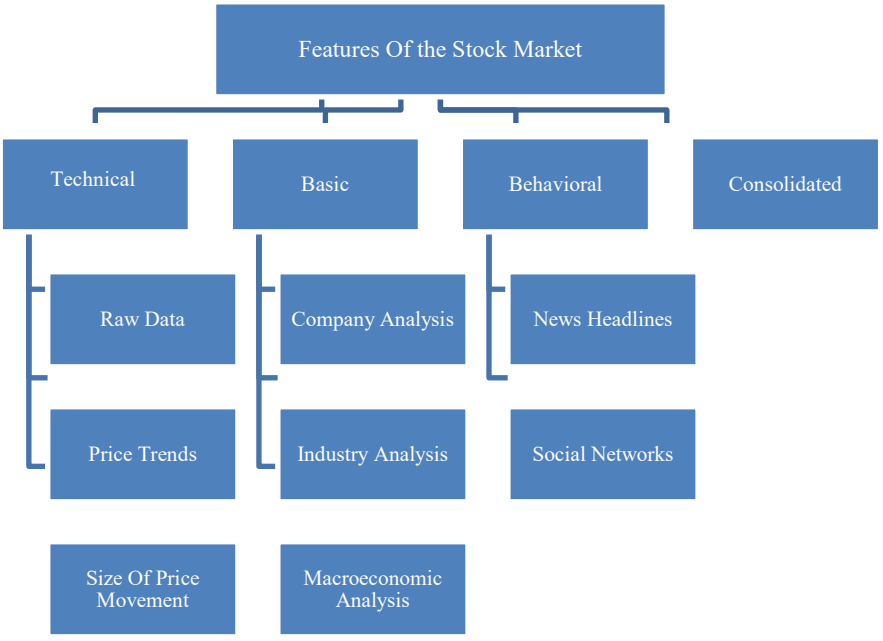

Fig. 2. Taxonomy of Features Used for Stock Market Prediction

4. CLASSIFICATION METHODS IN STOCK MARKET PREDICTION

In Section 3, various methods for extracting meaningful features were discussed. Based on those extracted features, different prediction models can now be proposed. Predicting the direction of stock price movements is essentially a classification problem, where the classifier's output can take three forms: +1, –1, and 0, representing an upward trend, a downward trend, and no change in stock prices, respectively.

Before the emergence of machine learning and artificial intelligence, researchers and investors primarily relied on mathematics-based approaches for stock prediction. Since these techniques were mainly capable of performing linear predictions, they became known as linear methods. In contrast, approaches based on machine learning owing to their ability to capture nonlinear patterns are generally referred to as nonlinear methods.

4.1. Linear Classification Methods for Stock Market Prediction

Najafabadi, in the Autoregressive (AR) model, states that the current value of a time series is a combination of one or more of its past values. This model represents the dependency of each value on its immediately preceding observations. The autoregressive process is a stochastic difference equation in which the current value is expressed as a function of previous values. The AR model includes a parameter P, which determines how many past observations are used to compute the current value [6].

A time series in a noisy environment is affected by random shocks; therefore, its current value is influenced by shocks that occurred in previous time points. The Moving Average (MA) model is designed to capture the effect of these random shocks on future values. The ARIMA model, which combines the autoregressive and moving average components, can approximate any stationary process to a desired degree of accuracy [7]. The 1990s marked the peak of ARIMA's popularity in time series forecasting, particularly in financial time series analysis. During this decade, the ARMA model was among the most widely used methods for financial forecasting.

However, Stock (1994) and Dunsmuir (1996) later identified a major limitation of these models related to the moving average component, a problem known as unit root accumulation. This issue occurs when the roots of the moving average polynomial become excessively large, causing the model's estimator to exhibit strong directional bias. Consequently, due to its simple linear structure and limited capacity for complex pattern recognition, this approach gradually lost favor among researchers [8].

The Box–Jenkins model, commonly referred to as ARIMA (Autoregressive Integrated Moving Average), was developed to improve upon previous methods. Although it offered a better predictive structure, ARIMA still shared the inherent limitations of this family of models namely, its reliance on historical values of a single variable to predict future outcomes. Hence, these models are known as univariate models. Since ARIMA cannot effectively capture nonlinear and complex relationships, its popularity among experts has significantly declined since the early 21st century [8].

Another widely used linear approach in time series forecasting is the Autoregressive Conditional Heteroskedasticity (ARCH) model, first introduced by Engle (1982) [9] and later extended by Bollerslev (1986) into the Generalized ARCH (GARCH) model [10]. The primary motivation for using ARCH-type models lies in their ability to model clusters of small and large forecast errors within a time series. The GARCH model describes dynamic changes in the conditional variance as a deterministic (often quadratic) function of past observations. Since the variance at time t–1 is known, a one-step-ahead forecast can be easily computed, and further steps can be recursively obtained [8]. Nevertheless, because this model represents changes through a low-order polynomial function and thus only partially captures nonlinearity, its performance in the highly complex and nonlinear stock market has proven limited.

As previously discussed, there is no universally accepted standard stock market prediction dataset for benchmarking various models. However, the accuracy results of linear and nonlinear methods are summarized in Tables 1 and 2, respectively. It should be noted that the performance evaluation metric used in the cited studies is the Root Mean Square Error (RMSE).

Table 1. Comparison of Accuracy in Linear Stock Market Prediction Methods

Error	Prediction Method	Feature Type	Reference
0.130	Auto Regressor	Technical	[6]
0.1235	ARIMA	Technical	[7]
0.143	ARMA	Technical	[8]
0.298	ARCH	Technical	[9]
0.282	GARCH	Technical	[10]

4.2. Nonlinear Classification Methods for Stock Market Prediction

Artificial intelligence (AI) methods, particularly optimization algorithms and machine learning techniques, have become some of the most widely used approaches for predicting financial time series. Examples include evolutionary algorithms, support vector machines (SVMs), fuzzy systems, and artificial neural networks (ANNs), all of which use past stock market data as input parameters and generate forecasts based on learned patterns. Among these, neural

networks constitute the core of most modern studies due to their capability to model nonlinear relationships and their robustness to noise.

In [11], a neural network was proposed that demonstrated effective real-time performance. Using Principal Component Analysis (PCA) during preprocessing, the model extracted the most informative features from the input data. When compared to backpropagation networks and SVMs, this model achieved superior performance. Yee (2015) [12] suggested enhancing neural networks through wavelet transformation. In [13], improvements were made to the neural network core by integrating it with a Petrie function within local linear models across hidden layers. In [14], a Functional Link Artificial Neural Network (FLANN) was employed, which can increase the dimensionality of input data. When properly executed, this higher-dimensional mapping can enhance network accuracy. The FLANN was implemented using a Multilayer Perceptron (MLP) structure.

In [15], the SVM model was optimized using both the Genetic Algorithm (GA) and Independent Component Analysis (ICA). The hybrid model SVM–GA achieved higher accuracy than both standard SVM and SVM–ICA. In [16], in addition to raw stock prices, technical indicators were used as network inputs. These indicators, derived from mathematical formulations, are widely applied in the literature, with MACD and RSI being among the most common. Incorporating these partially processed indicators enriches the input space and improves prediction accuracy. Since technical indicators are a logical and established means of constructing input vectors, a key question arises: Which indicators are the most effective? To address this, [17] combined neural networks with metaheuristic algorithms to identify the most relevant and accurate indicators.

Fundamental analysis has also been employed to improve network performance. For example, in [18], macroeconomic indicators such as oil prices, gold prices, and interest rates were used to identify the most influential factors affecting stock prices. The study concluded that oil price was the most significant among the fundamental indicators considered.

In [19], Japanese candlestick charts were incorporated alongside technical indicators as inputs to an RBF Neural Network (RBFNN). Since these charts capture information such as the daily high, low, opening, and closing prices as well as the bullish or bearish nature of market movements they provide highly informative input for prediction models.

In recent years, deep neural networks (DNNs) have gained considerable prominence. Deep learning is essentially an extension of neural networks with multiple hidden layers. As one moves deeper into the layers of a DNN, the model progressively learns more complex and abstract representations. Deep learning has been widely applied in various domains such as image classification, text categorization, speech recognition, and time-series forecasting, including stock market prediction.

In [20], a Convolutional Neural Network (CNN) was used to select optimal input features, while [21] employed a hybrid CNN–LSTM (Long Short-Term Memory) model to determine the most effective window size for stock prediction, where the window size refers to the number of past trading days considered in the forecast.

While CNNs are typically used for image-related tasks and have achieved remarkable results in that domain, [22] introduced an innovative idea: multiple chart types (e.g., candlestick, bar, and line charts) were plotted for each stock, and each chart was treated as an image input to train the CNN. The model then predicted the next image (t+1), corresponding to the future stock movement.

A separate line of research has focused on integrating behavioral and sentiment-based features into prediction models. In [24], Twitter sentiment analysis was conducted by classifying tweets about specific stocks, companies, or commodities as positive, negative, or neutral, and incorporating these sentiments into the prediction model alongside historical stock data. Since collective market sentiment can disrupt supply–demand equilibrium, incorporating such behavioral features improves prediction accuracy. A deep neural network was used for prediction in that study. Similarly, [25] utilized deep learning and sentiment analysis for stock forecasting, while [26] analyzed news headlines to extract sentiment polarity (positive, negative, or neutral). These headline-based sentiment features, whether derived from general news or stock-specific reports, were shown to enhance model precision.

In [23], LSTM networks were compared with SVMs, yielding competitive accuracy. In addition to closing prices, opening prices were also included in the calculations. Moreover, Naïve Bayes was used for sentiment analysis of

Twitter posts. In [22], a CNN was employed to explore dependencies among macroeconomic variables such as unemployment rate, interest rate, and liquidity ratio.

A more recent study, published in July 2019, combined sentiment analysis with stock price forecasting using an enhanced LSTM network. The proposed architecture introduced an attention layer to filter out less relevant information and emphasize critical data before feeding it into the LSTM. This attention-based LSTM architecture significantly improved prediction accuracy [27]. In that work, two modules were designed: one using a CNN to measure sentiment indices and another employing an attention-augmented LSTM for stock price forecasting.

Table 2 summarizes several representative nonlinear methods used in stock market prediction. As the results demonstrate, hybrid models, which leverage the strengths of multiple base techniques, typically achieve the best performance. Because of dataset inconsistencies and the absence of standard benchmarks, most studies compare their proposed models against baseline approaches such as MLP, SVM, or ARIMA.

Table 2. Comparison of Selected Nonlinear Methods for Stock Prediction

Error	Method	Feature Extraction	Source
0.1924	Principal Component Analysis + Time-Varying Neural Network	Technical analysis (raw data)	[11]
0.1672	Wavelet Transform + Neural Network initialized by Genetic Algorithm	Technical analysis (raw data)	[12]
0.0632	Neural Network based on Local Linear Radial Function	Technical analysis (raw data)	[13]
0.1593	Functional Link Neural Network	Technical analysis (raw data)	[14]
0.26	Support Vector Machine initialized by Genetic Algorithm	Technical analysis (raw data)	[15]
0.193	Fuzzy-Based Machine Learning Technique	Technical analysis (price trends)	[16]
0.1951	Hybrid of Meta-Heuristic Methods and Neural Network	Technical analysis (price trends)	[17]
0.23	Multilayer Perceptron Neural Network + Support Vector Machine	Technical + Behavioral + Fundamental analysis	[18]
0.2	Multilayer Perceptron Neural Network trained with KLD	Technical analysis (price trend + price momentum)	[19]
0.24	Deep Convolutional Neural Network + Wavelet Transform	Technical analysis (raw data)	[20]
0.226	Deep LSTM Neural Network	Technical analysis (raw data)	[21]
0.225	Deep Convolutional Neural Network + Feature Dependency	Technical analysis (price trends)	[22]
0.21	Deep LSTM Neural Network	Technical + Behavioral analysis	[23]
0.209	Deep Learning	Technical + Behavioral analysis	[24]
0.36	Deep Convolutional Neural Network	Technical + Behavioral analysis	[25]
0.18	Neural Network + Natural Language Processing Methods	Technical + Behavioral + Fundamental analysis	[26]
0.22	Deep LSTM Neural Network + Natural Language Processing Methods	Technical + Behavioral analysis	[27]

5. CONCLUSION

Today, the stock market has become a vital channel for mobilizing investment capital. The stock exchange is widely regarded as a benchmark for evaluating the financial and economic activities of a country or region. Predicting stock price movements remains one of the most challenging tasks in modern finance. Because the stock market is inherently dynamic, complex, nonlinear, and nonparametric, accurate forecasting of price fluctuations is crucial for developing effective trading strategies.

Researchers have employed a variety of methods to predict stock market behavior, leading to diverse outcomes. In this study, feature extraction methods were first reviewed across four main categories, followed by a detailed examination of classification models, including both linear and nonlinear approaches. Table 2 presents a comparative summary of recent artificial intelligence–based techniques used in stock market prediction.

As the table indicates, the error rates of most existing models remain relatively high, suggesting that there is still significant room for improvement. Studies that incorporated hybrid features combining technical, behavioral, and

fundamental analyses tended to achieve higher prediction accuracy. Given their intrinsic capability to model highly complex relationships, deep neural networks appear to be among the most promising solutions for addressing the challenges of stock price forecasting.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Jin, Z., Yang, Y., & Liu, Y. (2019). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications*, 1–17. https://doi.org/10.1007/s00521-019-04197-0

[2] Kim, T., & Kim, H. Y. (2019). Forecasting stock prices with a feature fusion LSTM–CNN model using different representations of the same data. *PLoS ONE, 14*(2), 1–23. https://doi.org/10.1371/journal.pone.0212320

[3] Rodrigues, A. A., & Lleo, S. (2018). Combining standard and behavioral portfolio theories: A practical and intuitive approach. *Quantitative Finance, 18*, 707–717. https://doi.org/10.1080/14697688.2017.1401225

[4] Abu-Mostafa, Y. S., & Atiya, A. F. (1996). Introduction to financial forecasting. *Applied Intelligence, 6*, 205–213. https://doi.org/10.1007/BF00126626

[5] Kim, K. J., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications, 19*, 125–132. https://doi.org/10.1016/S0957-4174(00)00027-0

[6] Najafabadi, S. R. M. (2009). *Prediction of stock market indices using machine learning* (Master's thesis). McGill University.

[7] Ababio, K. A. (2012). *Comparative study of stock price forecasting using ARIMA and ARIMAX models* (Master's thesis). Kwame Nkrumah University of Science and Technology.

[8] De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting, 22*(3), 443–473. https://doi.org/10.1016/j.ijforecast.2006.01.001

[9] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987–1007. https://doi.org/10.2307/1912773

[10] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

[11] Wang, J., & Wang, J. (2015). Forecasting stock market indexes using principal component analysis and stochastic time-effective neural networks. *Neurocomputing, 156*, 68–78. https://doi.org/10.1016/j.neucom.2014.12.084

[12] Ye, Q., & Wei, L. (2015). The prediction of stock price based on improved wavelet neural network. *Open Journal of Applied Sciences, 5*(4), 115–125. https://doi.org/10.4236/ojapps.2015.54012

[13] Patra, A., et al. (2017). An adaptive local linear optimized radial basis functional neural network model for financial time series prediction. *Neural Computing and Applications, 28*(1), 101–110. https://doi.org/10.1007/s00521-015-2039-0

[14] Gupta, A., Chaudhary, D. K., & Choudhury, T. (2017). Stock prediction using functional link artificial neural network (FLANN). In *Proceedings of the 3rd International Conference on Computational Intelligence and Networks (CINE)* (pp. 1–5). IEEE. https://doi.org/10.1109/CINE.2017.25

[15] Ahmadi, E., et al. (2018). New efficient hybrid candlestick technical analysis model for stock market timing on the basis of the support vector machine and heuristic algorithms of imperialist competition and genetic. *Expert Systems with Applications, 94*, 21–31. https://doi.org/10.1016/j.eswa.2017.10.023

[16] Patel, J., et al. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications, 42*(4), 2162–2172. https://doi.org/10.1016/j.eswa.2014.10.031

[17] Göçken, M., et al. (2016). Integrating metaheuristics and artificial neural networks for improved stock price prediction. *Expert Systems with Applications, 44*, 320–331. https://doi.org/10.1016/j.eswa.2015.09.029

[18] Usmani, M., et al. (2016). Stock market prediction using machine learning techniques. In *Proceedings of the 3rd International Conference on Computer and Information Sciences (ICCOINS)*. IEEE. https://doi.org/10.1109/ICCOINS.2016.7783235

[19] Hussein, A. S., Hamed, I. M., & Tolba, M. F. (2015). An efficient system for stock market prediction. In *Intelligent Systems* (pp. 871–882). Springer. https://doi.org/10.1007/978-3-319-11310-4_76

[20] Di Persio, L., & Honchar, O. (2016). Artificial neural networks architectures for stock price prediction: Comparisons and applications. *International Journal of Circuits, Systems and Signal Processing, 10*, 403–413.

[21] Selvin, S., et al. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. In *Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. https://doi.org/10.1109/ICACCI.2017.8126078

[22] Gunduz, H., Yaslan, Y., & Cataltepe, Z. (2017). Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. *Knowledge-Based Systems, 137*, 138–145. https://doi.org/10.1016/j.knosys.2017.09.023

[23] Li, J. H. B., & Wang, J. (2017). Sentiment-aware stock market prediction: A deep learning method. In *Proceedings of the International Conference on Service Systems and Service Management*. IEEE.

[24] Nivetha, R. Y., & Dhaya, C. (2017). Developing a prediction model for stock analysis. In *Proceedings of the International Conference on Technical Advancements in Computers and Communications (ICTACC)*. IEEE. https://doi.org/10.1109/ICTACC.2017.11

[25] Huang, Y., et al. (2016). Exploiting Twitter moods to boost financial trend prediction based on deep network models. In *Proceedings of the International Conference on Intelligent Computing* (pp. 679–688). Springer. https://doi.org/10.1007/978-3-319-42297-8_42

[26] Attigeri, G., et al. (2015). Stock market prediction: A big data approach. In *Proceedings of TENCON 2015—IEEE Region 10 Conference* (pp. 1–6). IEEE. https://doi.org/10.1109/TENCON.2015.7373006

[27] Jin, Z., Yang, Y., & Liu, Y. (2019). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications*, 1–17.