A hybrid sentiment based stock price prediction model using machine learning

Awais Mehmood[1*], and *Muhammad* Khurram Ali[1]

[1*, 1] Department of Industrial Engineering, University of Engineering and Technology Taxila, Punjab 47050, Pakistan

Abstract. Accurate stock market prediction is highly desirable to corporations and investors. In this study a deep learning model based on LSTM, BiLSTM with attention mechanism used to predict stocks closing price for next 30 days of two banks listed in Pakistan Stock Exchange. For accurate stock price prediction, it is necessary to consider volatile factors such as news sentiments along with historical data. This study covers that aspect by incorporating news sentiments along with historical stock data that is distributed over a span of ten years from Jan 2011 to July 2021. Preprocessing and sentiment analysis of data was performed using python NLTK module. After that we built a univariate deep learning model based on four layers of LSTM and one dense layer to combine all layers and performed a prediction on train and test data followed by a multivariate deep learning model based on BiLSTM with self-attention mechanism and found out that incorporation of news sentiments really improved the prediction accuracy by reducing the values of mean squared error. Finally, we did the prediction for next 30 days of stock closing price of two banks and compared those predicted prices with actual prices and got quite accurate results.

Keywords: Deep Learning, Stock price prediction, News Sentiments, LSTM, BiLSTM

1 Introduction

A country's economic development is greatly influenced by the stock market. The introduction of internet trading has fundamentally revolutionized the way individuals purchase and sell stocks. Financial technology is a brand-new financial industry dedicated to enhancing financial market activity via the use of cutting-edge technological solutions [1]. When it comes to the supply of financial services, fintech is posing a threat to the status quo. Things occurred swiftly in the global financial markets. As long as the stock market forecaster is accurate, investors may expect little risk and a substantial profit.

*Corresponding author: owaisniazi94@gmail.com

One of the most difficult and time-consuming tasks of the past several decades has been making accurate predictions about the stock market. Researchers from several domains have experimented with various Machine Learning, Deep Learning, and statistical approaches for financial time-series forecasting. Stock market history and predicting financial news sentiments are two methods of doing this. These forecasting approaches keep an eye on, control, and predict the stock market's value to maximize gains while reducing dangers. However, researchers are still searching for a satisfactory solution to the challenge of resilient, efficient, and accurate prediction. Stock market forecasting is hindered by the non-parametric and stochastic nature of financial time series data [2]. In time series or stock forecasting models, linear regression and moving averages are standard tools. In the past, regression models such as auto-regression, auto-regression moving average, auto-regression integrated moving average, and support vector regression (SVR) were often used [3]. In addition to ARIMA with Explanatory Variables and SARIMA, researchers have devised several different variants of these regression models throughout the years. A stock's future price may be predicted using machine learning and deep learning-based approaches, which uses both numerical and textual data in order to consider the most recent information.

Many firms use linear regression models like ARMA, ARIMA, ARIMAX, and SARIMA, all of them have a performance lag. Using a stock data prediction model that works well for one company may not perform well for another [4].

Some of the most often used models for predicting stock values are autoregressive moving average, autoregressive integrated moving average, support vector regression, support vector machine, random forest, and artificial neural networks [5]. Since the stock market's trajectory is influenced by so much noise, non-linearity, stochastic and chaotic nature of data, in that scenario successful stock price forecasting may lead to large gains and lower risk. Because of recent developments, including frameworks for stock market forecasting, traders and investors are paying more attention to the stock market. Although stock market forecasting may be challenging, scientists are striving to develop algorithms that might help predict the market's future path. Investors may benefit from utilizing the models provided to forecast and monitor the stock market [6]. The stock market's state may be affected by a variety of factors, including financial news, comments from financial experts, stock market content on social media, the annual budget release, and financial reports. The use of precise stock prediction algorithms helps traders and investors make better buying and selling decisions. Superior stock market data analysis and extraction may have a significant impact on tools that help investors predict future trends and behaviors [7]. Data mining and knowledge discovery from databases have both been proposed for stock market analysis [8].

2 Related Work

SVM and ANN were analyzed by [9] to predict price fluctuations in the stock market. Three layered feed-forward-NNs are designed for input and output. In terms of both structure-complexity and danger of optimal trade-off, the SVM merges with learning theory. The SVM hyperplane expresses positive and negative data. The accuracy for both Neural Network and SVM approaches were 75.73 percent and 71.53 percent, respectively. For critical variables, there are no rules for setting the parameters, and both models have detailed parameters.

Artificial Neural Network, SVM, Random Forest, and Naïve-Bayes for accurate estimation was adopted by [10]. In this work, the investigator used Two-Approaches for the review of these models. In the first system, ten technical metrics (open price, close price, low price and high price,) were calculated using these models as an input parameter. The probabilistic data preparation-layer strategy was used in the second approach for the transformation of continuous data to discrete data by applying the Up / down comparison

indexes of the stock market. To validate this study, the researchers used historical data and compared the results of these models. It is noted that the suggested input-approach shows-best-result in the sense of reliability. The accuracy was 86.6 percent in Artificial Neural Network, 89.34 percent in SVM, 89.89 percent in Random Forest and 90 percent in Naïve-Bayes.

In [11] two strategies were proposed for the prediction of future stock-market index value. In this research work, two distinct layers were used; support-vector regression (SVR) was used to predict future significance in the first layer. The future value was used as an input in the second layer and is combined with the ANN, Random Forest and SVM for the forecast models. The consequence of the suggested method was a contrast to a single-stage solution. It is noted that the suggested approach demonstrates better performance in the sense of precision. This improved the reliability of 11.5 percent of ANN vs. SVR-ANN, 1.66 percent of SVR vs. SVR-SVR and 9.1 percent of RF vs. SVR-RF. The proposed model is accurate and stable, it can reduce efficiency by increasing parameters.

A hybrid model was proposed by [12] named as Levenberg Marquardt neural-network (PELMNN) based on adaptive pre-processing for the prediction of stock market. In the beginning, Genetic Algorithm (GA) was used for determining the optimal weight of the artificial-neural network (ANN). To minimize Levenberg Marquardt-BP entry, the preprocessing approach was used twice. To verify 50 price indices, the investigator used PELMNN. The results of the proposed strategies were noted, compared the hybrid fuzzy model and ANN, the proposed approach achieves better efficiency and estimation accuracy having a value of 0.51 percent while the hybrid fuzzy model's and ANN MAPE were 1.3 percent and 0.78 percent respectively.

A synthetic approach was adopted by [13] to predict the Japanese stock market. As an input for maps of non-linear data in this method, new data collection was introduced. The classical-back-propagation learning-algorithm has been used in this step for efficient return. In this method, GA and simulated annealing (SA) were used to boost forecasting while preprocessing was used to minimize search space and fuzzy-curve checking to determine similarity in the input-output parameter. The investigator used historical knowledge to validate this method. The consequence of the proposed model was a reference to BP-NN. It is remembered that the proposed model shows the better accuracy principle and minimizes errors. In the best case, the efficiency was 0.0725 using 28 Processor times while using 68 BB-NN, CPU time was 0.0044.

In [14], the basic model of ANN was implemented to forecast the financial market. This framework is the synthesis of science and fundamental study of economic and financial principles using time series. These technical and fundamental studies have been used to predict the potential behavior of stock prices.

Random Forest was used by [15] to predict stock price based on social media, financial news and stock exchange data with the prediction accuracy of 83%. The dataset comprises of previous 2 years. Sentimental analysis was also performed to get useful features from news. This study shows that Hybrid approaches works better than other approaches.

In [16] Artificial Neural Networks was used for prediction of stock prices based on both historical data and news sentiments data distributed over a span of five years from 1st Jan 2015 to 31 Dec 2020.

Artificial Neural Network and Linear regression were adopted by [17] for prediction of stock closing price of five different companies belonging to different sectors listed in Karachi stock exchange. Last ten years of historical data was used to predict next day stock closing price and found out that ANN model performed better. For future they suggested to include news data along with deep learning models to improve prediction accuracy.

Predictive analysis on markets of four countries US, Hong Kong, Turkey and Pakistan was performed by [18] . Sentimental analysis is performed using 4 years of twitter data. It is

found out that deep learning techniques outperformed Linear Regression and Support Vector Regression.

3 Methodology

Now we will discuss, the detailed methodology of proposed Machine learning Model to predict stock closing prices. Initially the techniques used for stock price prediction and the evaluation criteria have been identified from literature, then acquired the historical stocks data along with the news data, distributed over a span of 10 years for two banks. After that we performed a univariate analysis based on LSTM followed by a multivariate analysis based on BiLSTM with attention mechanism. Then we evaluated the model's performance based on Mean Squared Error. Finally predicted the next 30 days stocks closing price for both banks. Step by step proceedings of this study are explained below:

3.1 Stocks Data Collection

From January 2011 to July 2021, we gathered stock data of two banks. Data is saved in the form of .csv file so that it can easily be imported and analyzed in the python.

3.2 News data Collection

News data acquired from credible source like Dawn news for two banks.

3.3 Preprocessing of News data

The acquired news data contains a lot of irrelevant characters, symbols, lower/upper case and other redundant information which needs to be removed before feeding to the proposed deep learning model. Following preprocessing steps were performed:

3.3.1 Data cleaning

/, @, A, #, *, etc. are some of the useless symbols seen in the raw data of the news. To proceed with the processing, these few useless symbols must be eliminated. In the preparation stage, certain symbols, punctuation, and superfluous spaces were deleted.

3.3.2 URL/hyperlinks Removal

In order to analyze the data, all URLs and linkages have been deleted. So, we don't need to examine URLs unless it is necessary for our analysis. Regular expressions (RE) was used to eliminate all URLs.

3.3.3 Case-folding

Capital letters are not required in our scenario. So, all uppercase letters should be converted to lowercase. "United" and "united "are considered as two separate words if case-folding is not performed.

3.4 Normalization

Stocks data and news sentiment were normalized using min-max normalization using the following procedure.

$$\frac{\text{value-min}}{\text{max-min}} \tag{1}$$

Min-max normalization is a simple and widely used approach for data normalization. The lowest and highest values for each feature should be set to 0 and 1, respectively, while the remainder of the values should stay between 0 and 1.

3.5 Prediction Model

After all the preprocessing of data we distributed it into train and test data by the ratio of 80:20 (80% training data and 20% test data). After that we applied a univariate deep learning model based on LSTM followed by a multivariate model based on BiLSTM with attention mechanism to improve the prediction accuracy by reducing the error.

3.6 Evaluation

The evaluation criteria we used is the Mean Squared Error (MSE), which is the mean difference between the actual and predicted values.

$$\text{MSE}=\frac{\sum_{i=1}^{n}(\text{Predicted}_i\text{-Actual}_i)^2}{n} \tag{2}$$

4 Results and discussion

Now we will discuss the results we acquired after applying univariate and multivariate deep learning models on datasets of two banks to predict their next 30 days stock closing prices.

4.1 Results of Univariate and Multivariate Analysis on Bank 1 Dataset

Figure 1 and 2 given below shows that our predictions are quite accurate as our predicted values curve nicely follows the actual values curve.

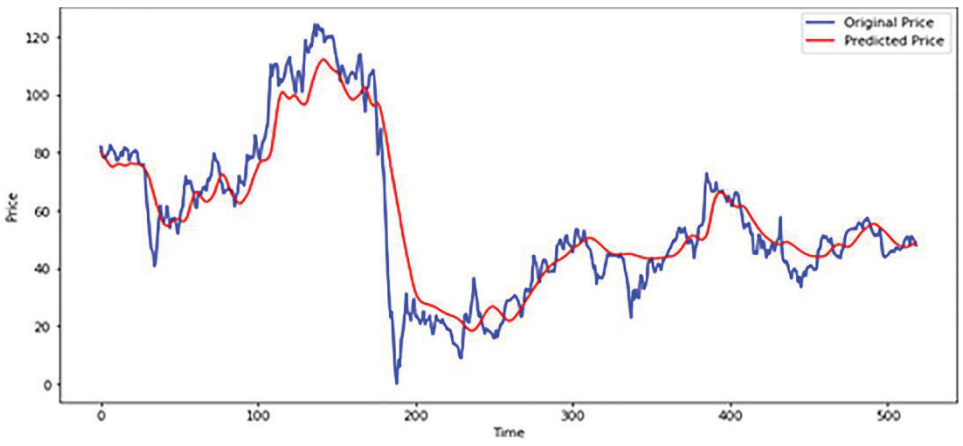

Fig. 1. Univariate Analysis on Bank 1 Dataset

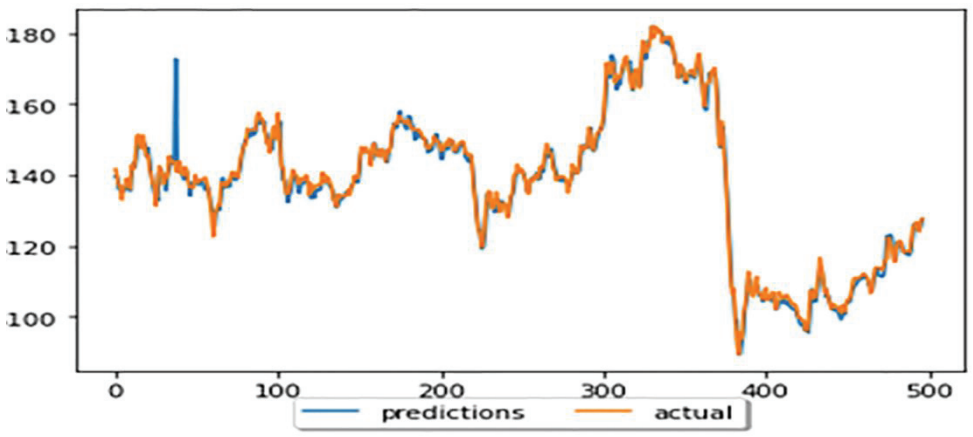

Fig. 2. Multivariate Analysis on Bank 1 Dataset

4.1.1 Model Evaluation

Now we will evaluate both models based on Mean Squared Error.

Table 1. MSE Scores of Univariate and Multivariate Models for Bank 1 Dataset

Mean Squared Error (MSE)	**Univariate Model**	**Multivariate Model**
	0.0075	0.0010

The above table shows that after incorporating news and applying BiLSTM based deep learning model prediction accuracy get increased.

4.1.2 Predicting Next 30 days Stock Closing Price of Bank 1

As our Machine learning model is now trained and tested. Now we can calculate next 30 days stock closing price for bank 1. Finally, we compared those predicted values with actual values and calculated the difference in the form of percentage.

Table 2. Bank 1 Stock Closing Price Comparison

Sr No	**Date**	**Actual Values**	**Predicted Values**	**Percentage Difference**
1	8/2/2021	125.33	126.107139	-1%
2	8/3/2021	125.23	126.2508446	-1%
3	8/4/2021	126.49	126.3460821	0%
4	8/5/2021	126.85	126.4085186	0%
5	8/6/2021	127.21	126.4521889	1%
6	8/9/2021	127.47	126.487306	1%

7	8/10/2021	127.69	126.5201649	1%
8	8/11/2021	128.67	126.553953	2%
9	8/12/2021	129.2	126.5897898	2%
10	8/13/2021	129.51	126.6276421	2%
11	8/16/2021	129	126.6669641	2%
12	8/17/2021	129.36	126.7071023	2%
13	8/20/2021	129.33	126.7474914	2%
14	8/23/2021	129.7	126.7877233	2%
15	8/24/2021	129.48	126.8275334	2%
16	8/25/2021	129.01	126.8667864	2%
17	8/26/2021	124.28	126.905568	-2%
18	8/27/2021	123.07	126.9439167	-3%
19	8/30/2021	122.19	126.9817277	-4%
20	8/31/2021	120.93	127.0189597	-5%
21	9/1/2021	121.53	127.0556043	-5%
22	9/2/2021	117.97	127.0916534	-8%
23	9/3/2021	117.65	127.1271262	-8%
24	9/6/2021	116.12	127.16202	-10%
25	9/7/2021	115.58	127.1963485	-10%
26	9/8/2021	114.37	127.2301201	-11%
27	9/9/2021	115.97	127.2633346	-10%
28	9/10/2021	115.98	127.2960005	-10%
29	9/13/2021	121.69	127.3281314	-5%
30	9/14/2021	120.75	127.3597247	-5%

4.2 Results of Univariate and Multivariate Analysis on Bank 2 Dataset

Figure 3 and 4 given below shows that our predictions are quite accurate as our predicted values curve nicely follows the actual values curve.

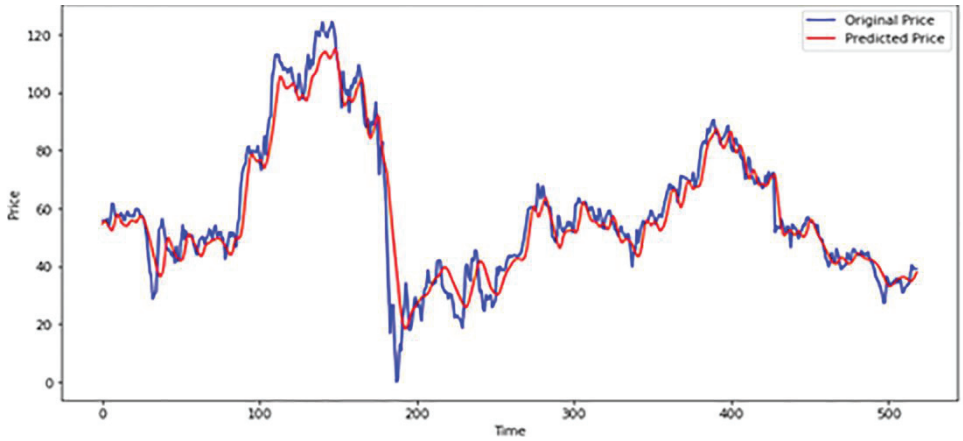

Fig. 3. Univariate Analysis on Bank 2 Dataset

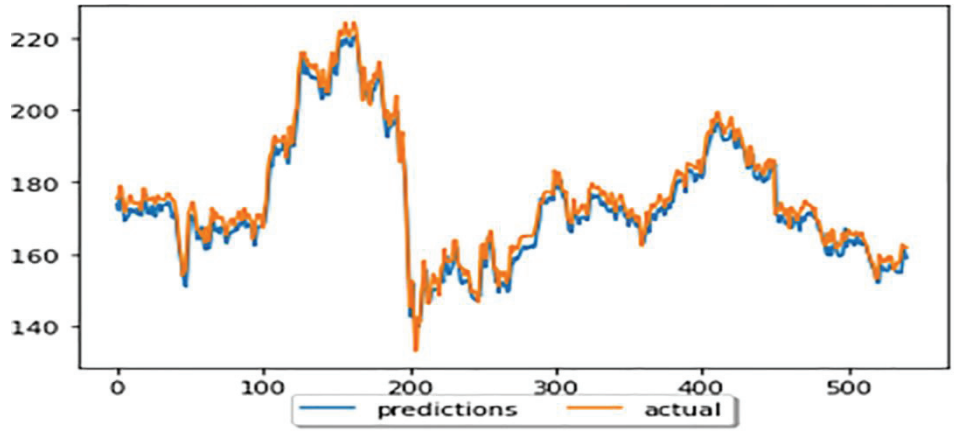

Fig. 4. Multivariate Analysis on Bank 2 Dataset

4.2.1 Model Evaluation

Now we will evaluate both models based on Mean Squared Error.

Table 3. MSE Scores of Univariate and Multivariate Models for Bank 2 Dataset

Mean Squared Error (MSE)	**Univariate Model**	**Multivariate Model**
	0.0034	0.0015

The above table shows that after incorporating news and applying BiLSTM based deep learning model prediction accuracy get increased.

4.2.2 Predicting Next 30 days Stock Closing Price of Bank 2

As our Machine learning model is now trained and tested. Now we can calculate next 30 days stock closing price for bank 2. Finally, we compared those predicted values with actual values and calculated the difference in the form of percentage.

Table 4. Bank 2 Stock Closing Price Comparison

Sr No	Date	Actual Values	Predicted Values	Percentage Difference
1	8/2/2021	164.49	161.5236519	2%
2	8/3/2021	166.11	162.0045195	2%
3	8/4/2021	166.02	162.2467612	2%
4	8/5/2021	165.37	162.3289368	2%
5	8/6/2021	166.12	162.3288904	2%
6	8/9/2021	167.43	162.2991926	3%
7	8/10/2021	165.58	162.2664524	2%
8	8/11/2021	164.7	162.2367545	1%
9	8/12/2021	163.98	162.2070868	1%
10	8/13/2021	163	162.1752833	1%
11	8/16/2021	161.84	162.1406559	0%
12	8/17/2021	163.5	162.1042041	1%
13	8/20/2021	164.91	162.0681374	2%
14	8/23/2021	167.01	162.0348209	3%
15	8/24/2021	165.57	162.0059861	2%
16	8/25/2021	166.69	161.9822666	3%
17	8/26/2021	168.04	161.9637635	4%
18	8/27/2021	166.87	161.9500534	3%
19	8/30/2021	165.48	161.9404591	2%
20	8/31/2021	165.25	161.9341722	2%
21	9/1/2021	163.72	161.9304306	1%
22	9/2/2021	160.6	161.9285953	-1%
23	9/3/2021	161.95	161.9281256	0%
24	9/6/2021	162.82	161.9285735	1%
25	9/7/2021	163.09	161.9295649	1%
26	9/8/2021	162.9	161.9307966	1%

27	9/9/2021	157.71	161.9320365	-3%
28	9/10/2021	156.04	161.9331125	-4%
29	9/13/2021	157.04	161.9339264	-3%
30	9/14/2021	155.58	161.9344262	-4%

10 Conclusion

This research proposes a new taxonomy based on widely used approaches for accurate stock market forecasts. In literature review we found out that stock price prediction is performed either using conventional learning techniques such as Support Vector machines, Artificial Neural Networks. Logistic Regression and K-nearest neighbors or deep learning techniques such as Recurrent Neural Networks, Convolutional Neural Networks and Long Short Term Memory, but found out that hybrid models such as LSTM with self-attention mechanism outperforms all other models in terms of prediction accuracy. Each examined study's research gaps and future prospects in the area of stock price prediction are discussed in detail, and this is the most significant contribution of the work. Considering the efficacy of hybrid model this study is also based on that approach.

The dataset we used for both banks consist of 2594 rows which is quite comprehensive amount of data for analysis. The performance of univariate model based on LSTM for both banks test data was very good resulting in MSE of 0.0075 and 0.0034 respectively. Then we applied a multivariate model based on BiLSTM with self-attention mechanism on both banks dataset which resulted in MSE value of 0.0010 and 0.0015 respectively. Hence by incorporating news prediction accuracy get improved. After that we predicted the stock closing price for next 30 days and got good results. The difference between actual and predicted prices for both banks were in the range of 0 to 11% and 0 to 4% respectively. So, we can suggest that investors can rely on this study while making investment decision in banking sector.

As the future directions and limitation of the work that need to be addressed to improve stock prediction accuracy. The BERT (Bidirectional Encoder Representations from Transformers) is a high-performance natural language processing model announced by Google, that can be used to construct a more powerful and accurate prediction model for stock prices. BERT has the potential to provide new information and to improve forecast accuracy.

References

[1] T. Puschmann, "Fintech," *Business & Information Systems Engineering,* vol. **59**, no. 1, pp. 69-76, 2017

[2] M.-C. Tsai, C.-H. Cheng, M.-I. Tsai, and H.-Y. Shiu, "Forecasting leading industry stock prices based on a hybrid time-series forecast model," *PloS one,* vol. **13**, no. 12, p. e0209922, 2018

[3] D. Bhuriya, G. Kaushal, A. Sharma, and U. Singh, "Stock market predication using a linear regression," in *2017 international conference of electronics, communication and aerospace technology (ICECA)*, 2017, vol. **2:** IEEE, pp. 510-513.

[4] M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," *Applied soft computing,* vol. **11**, no. 2, pp. 2664-2675, 2011

[5] K. H. Sadia, A. Sharma, A. Paul, S. Padhi, and S. Sanyal, "Stock market prediction using machine learning algorithms," *Int. J. Eng. Adv. Technol,* vol. **8**, no. 4, pp. 25-31, 2019

[6] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[7] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, "Big Data: Deep Learning for financial sentiment analysis," *Journal of Big Data,* vol. **5**, no. 1, pp. 1-25, 2018

[8] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction using the ARIMA model," in *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, 2014: IEEE, pp. 106-112.

[9] Y. Kara, M. A. Boyacioglu, and Ö. K. J. E. s. w. A. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange," vol. **38**, no. 5, pp. 5311-5319, 2011

[10] J. Patel, S. Shah, P. Thakkar, and K. J. E. s. w. a. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," vol. **42**, no. 1, pp. 259-268, 2015

[11] J. Patel, S. Shah, P. Thakkar, and K. J. E. S. w. A. Kotecha, "Predicting stock market index using fusion of machine learning techniques," vol. **42**, no. 4, pp. 2162-2172, 2015

[12] S. Asadi, E. Hadavandi, F. Mehmanpazir, and M. M. J. K.-B. S. Nakhostin, "Hybridization of evolutionary Levenberg–Marquardt neural networks and data pre-processing for stock market prediction," vol. **35**, pp. 245-258, 2012

[13] M. Qiu, Y. Song, F. J. C. Akagi, Solitons, and Fractals, "Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market," vol. **85,** pp. 1-7, 2016

[14] F. A. de Oliveira, C. N. Nobre, and L. E. J. E. S. w. A. Zarate, "Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index–Case study of PETR4, Petrobras, Brazil," vol. **40**, no. 18, pp. 7596-7606, 2013

[15] W. Khan *et al.*, "Stock market prediction using machine learning classifiers and social media, news," pp. 1-24, 2020

[16] A. Zaffar and S. M. A. Hussain, "Modeling and prediction of KSE – 100 index closing based on news sentiments: an applications of machine learning model and ARMA (p, q) model," *Multimedia Tools and Applications,* vol. **81**, no. 23, pp. 33311-33333, 2022/09/01 2022

[17] M. Hameed, K. Iqbal, R. Ghazali, F. H. Jaskani, and Z. Saman, "Karachi Stock Exchange Price Prediction using Machine Learning Regression Techniques," *EAI Endorsed Transactions on Creative Technologies,* vol. **8**, no. 28, p. e5, 08/24 2021

[18] H. Maqsood *et al.*, "A local and global event sentiment based efficient stock exchange forecasting using deep learning," vol. **50**, pp. 432-451, 2020