International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 04 | Apr 2024 www.irjet.net p-ISSN: 2395-0072

# STOCK MARKET PREDICTION USING HYBRID LSTM-ARIMA AND SENTIMENT ANALYSIS ON NIFTY STOCK INDEX

**Dr. S. Annie Joice[1], U. Mohamed Baashith[2], R. Manikandan[3], P. Sathasiva Pandi[4]**

*[1]Assistant Professor, Department of CSE, Government College of Engineering, Srirangam, Tamilnadu, India*
*[2,3,4]UG student, Department of CSE, Government College of Engineering, Srirangam, Tamilnadu, India*

---

**Abstract -** *This paper presents a hybrid approach for stock market price prediction by integrating Long Short-Term Memory (LSTM), Auto Regressive Integrated Moving Average (ARIMA), and sentiment analysis derived from Twitter data. The proposed hybrid LSTM-ARIMA model involves gathering company-specific data, segmenting it into three components based on trend, seasonal, and residual characteristics, and incorporating tweets for sentiment analysis. The proposed model utilizes LSTM to handle seasonal and residual data, ARIMA for trend data, and sentiment analysis to gauge market sentiment. The hybrid model demonstrates enhanced predictive accuracy compared to standalone LSTM and ARIMA models, underscoring the efficacy of integrating these methods. Tweets obtained from Twitter are utilized to inform decisions regarding stock transactions. Results highlight the significant impact of sentiment analysis on improving overall prediction performance. The hybrid LSTM-ARIMA model outperforms the standalone models and gives an MAE of 0.58, RMSE of 1.16, MAPE of 0.01. The hybrid LSTM-ARIMA model achieves best accuracy for every stock when contrasted with the two standalone models (ARIMA & LSTM). This paper contributes to the advancement of stock market prediction methodologies and recommends the user to buy or sell a particular stock.*

***Keywords***: **Hybrid LSTM-ARIMA Model, Sentiment Analysis, Seasonal, Residual, Trend**

## 1. INTRODUCTION

The stock market, as a dynamic and complex system, constantly challenges investors and analysts to develop more accurate prediction models. In recent years, machine learning techniques [1] have emerged as powerful tools for forecasting financial markets.

Historical stock market prediction methods have predominantly relied on singular models, such as ARIMA or LSTM [2] [3] [4]. This paper aims to bridge the gap between LSTM and ARIMA, two widely used models, by combining them to better understand the complex patterns in financial data. LSTM excels at capturing long-term dependencies, while ARIMA is effective in modeling trend data. This combination leverages the strengths of each model, creating a synergy that promises enhanced forecasting accuracy. In addition to the hybrid modeling approach [5], this paper introduces sentiment analysis as a crucial component [6]. In an era dominated by social media, understanding the impact of public perception on stock prices is paramount. By incorporating sentiment analysis of tweets related to a specific company, the model integrates real-time market sentiment, providing a more holistic perspective. The choice of data sources is pivotal to any prediction model. For the hybrid approach, historical stock market data is sourced, which is then divided into seasonal, residual, and trend components. Concurrently, relevant tweets from Twitter pertaining to the selected company are collected, providing the sentiment analysis algorithm with a rich dataset. This meticulous curation of diverse data sources ensures that the model is not only comprehensive but also capable of adapting to the multifaceted nature of stock market dynamics. This paper is driven by the recognition that accurate stock market predictions empower investors to make informed decisions, mitigating risks and maximizing returns. By embracing a hybrid LSTM & ARIMA model with sentiment analysis, this paper aspires to contribute to this evolving landscape, pushing the boundaries of what is possible in forecasting stock market prices.

This paper investigates into this realm, proposing a hybrid approach [5] that combines LSTM and ARIMA models with sentiment analysis [6]. This integration aims to overcome the limitations of stand-alone models, providing a more robust and reliable prediction framework.

The rest of the paper is organized as follows: Section 2 summarizes the related work on stock market price prediction. Section 3 describes the proposed hybrid LSTM-ARIMA model with sentiment analysis, as well as the training procedure. Section 4 describes the dataset, and its preprocessing for training. Further, the experimental results and performance analysis are also included in this section. In Section 5, the conclusion and future research direction are highlighted.

## 2. RELATED WORK

Researchers and practitioners extensively investigated various approaches to develop models capable of forecasting stock prices, addressing the challenging and complex task of stock market price prediction. In their studies, they explored key areas and methods commonly used in this field. Zhao et al. [7] employed time series relational models for stock price prediction, enhancing

forecasting accuracy by analyzing historical data and incorporating relational structures. Widiputra et al. [8] proposed a model that combined Convolutional Neural Networks (CNN) and LSTM networks to predict multiple parallel financial time series, leveraging CNN for feature extraction and LSTM for sequential learning. Nelson et al. [9] trained and evaluated an LSTM model for predicting an increase in the stock's price within the next 15 minutes. Bhandari et al. [10] proposed a model implemented in single and multi-layer LSTM architectures, and their performance was analyzed using various evaluation metrics to identify the best model. Hatano et al. [11] proposed a model that utilized features generated using Sentence-BERT for tweet data to train LightGBM (Gradient Boosting Machine), to predict sudden changes in closing prices. Antad et al. [12] proposed a model focused on using historical data for prediction and aimed to enhance precision by employing linear regression models. Kapoor et al. [13] proposed a model that utilized multi-layer hierarchical LSTM networks to predict stock prices. Leveraging deep learning techniques, LSTMs processed historical data to forecast future prices. This necessitated computational resources and programming expertise in a machine learning framework such as TensorFlow or PyTorch. Kaeley et al. [14] proposed a model tested for longer time windows and incorporated new technical indicator features to improve the performance of RNN models in stock price prediction. Aldhyani et al. [15] proposed a hybrid model that integrates Convolutional Neural Network (CNN) and LSTM networks to forecast the closing prices of companies using deep learning techniques. Alfonso et al. [16] illustrated with an application to the stock market using neural networks to predict the stock prices with less error rates. Cao et al. [17] introduced a hybrid forecasting model combining LSTM and CEEMDAN, with the goal of improving the accuracy of stock market price predictions. The performance of the model is assessed through linear regression analysis. Feng et al. [18] introduced a deep learning technique called Relational Stock Ranking (RSR) to rank stocks and capture their relationships in a time-sensitive manner. Wang et al. [19] introduced a novel method termed Adaptive Long-Short Pattern Transformer (ALSP-TF) for stock ranking, demonstrating superior experimental results compared to state-of-the-art stock forecasting techniques. Zhao et al. [20] introduced a unified time-series relational multi-factor model (TRMF), incorporating a self-generating relations (SGR) algorithm capable of automatically extracting relational features.

## 3. PROPOSED SYSTEM

### 3.1 System Architecture

Leveraging the integration with Yahoo Finance, historical stock price details for the chosen company over the past years are fetched seamlessly. Simultaneously, recent tweets related to the company are pulled from Twitter to provide additional contextual information. These tweets undergo Natural Language Processing (NLP) to prepare them for sentiment analysis, discerning whether they convey positive, negative, or neutral sentiments [17]. The fetched stock market data undergoes preprocessing to enhance accuracy. Historical stock market data is decomposed into trend, seasonal, and residual components. Each of the components is trained on their own. Prediction results from all three components were added for the final prediction of stock.

Subsequently, a Hybrid LSTM & ARIMA model is employed to forecast the future stock price of the chosen company. Once trained, the model generates forecasts based on the learned patterns from the historical stock price data and sentiment analysis of relevant tweets. Upon completion of the training and testing phases, the algorithm provides a forecast value for the stock price of the particular company. Figure 1 demonstrates the overall system architecture of the stock market price prediction model.

The proposed system is divided into three major components: a) Collection of the stock data such as the open, close, high, low, adjusted close prices & volume and preprocessing of tweets for sentiment analysis, b) classification of stock ticker using LSTM, ARIMA and Hybrid LSTM-ARIMA and c) Performance Analysis of the different models.

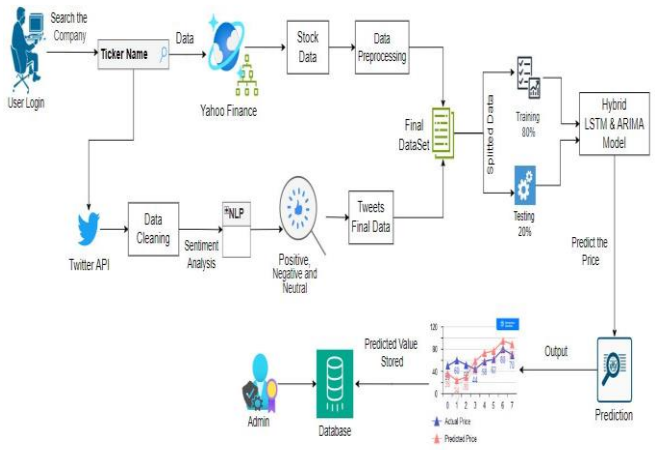



**Fig -1**: System Model

### 3.2 Methodology

#### 3.2.1 Collection of Stock Ticker

Dataset is prepared from tickers of historical stock market data from Yahoo Finance API and the tweets for the sentiment analysis are extracted using Twitter API. To conduct the study, data from the Indian Stock Market's National Stock Exchange (NSE) Nifty 50 index spanning from Jan 1, 2022 to the current date, encompassing nearly 2 years of data. Latest tweets for the past 7 days are fetched to get

accurate prediction results. The implementation of this model ensures a more effective discernment of sentiments expressed in tweets, ultimately assisting in assessing community perceptions of the company and its potential impact on the stock market.

### 3.2.2 Long Short-Term Memory (LSTM)

LSTM is an efficient method for time series forecasting. LSTM can be able to remember information over a long period of time and it is best suited for stock market price prediction [6] [7] [9]. It helps us understand patterns over time. It looks at the seasonal and leftover (residual) parts of the stock data. This approach trains the LSTM on some data and then checks how well it does on new data. This way, it learns from the past and applies that learning to make predictions. LSTM is like a sophisticated tool that remembers and understands how the stock market behaves over different seasons, helping to make better prediction. LSTM is a type of recurrent neural network (RNN) designed to model temporal sequences and long-term dependencies. In the context of stock market prediction, LSTM can be used to learn patterns and trends from historical stock price data.

### 3.2.3 Auto-Regressive Integrated Moving Average (ARIMA)

The Auto-Regressive Integrated Moving Average (ARIMA) method focuses on predicting stock prices by understanding and analyzing the long-term trend in the data. Trained on historical data representing the trend component, ARIMA provides insights into sustained market movements. After training, the model's effectiveness is assessed using a separate testing dataset. This step ensures that ARIMA's capacity to capture and forecast trends in stock prices is accurate and reliable. By concentrating on the underlying trends in the stock market data, ARIMA complements the short-term focus of other models, contributing to a more comprehensive and robust prediction framework. The ARIMA model is a popular time series forecasting method, commonly used for stock market price prediction.

### 3.2.4 Sentiment Analysis Model

In this paper, the tweets about the selected company have been extracted from Twitter. Tweets not relevant to the company are removed. Text preprocessing of tweets are performed using Regular Expressions. The presence of hashtags and specifies in the text increases the length of the message and thereby decreasing the ability of the analysis model to classify the sentiments. Hence, they have been removed. Further, the tweets have been converted to lowercase and multiple whitespace characters have been replaced with a single whitespace character. After preprocessing the dataset additional tree features will be used to highlight the features which are the most effective to predict the closing price of the stock. Scikit-learn implement an estimator that uses randomized decision tree to fit different subset of the dataset by utilizing the mean method. Hence the prediction accuracy of the model is improved and the overfitting of the dataset is controlled. For our study the proposed model extracts all tweets that specify either the company name, stock name or name of any of the board members currently in the company.

### 3.2.5 Hybrid LSTM and ARIMA Model

Each of the models ARIMA and LSTM has their own strengths and weaknesses. ARIMA does not work well with non-linear timeseries. On the other hand, LSTM can work with linear and non-linear time series data. However, LSTM requires long training time. By considering these factors, a hybrid model is implemented and it works with their own expertise. The proposed hybrid model is expected to give more accurate predictions compared to the standalone models. The hybrid model can make models work together to overcome each other's weaknesses. The hybrid model captures both short-term patterns and long-term trends, providing a more comprehensive and subtlety prediction of stock prices. The synergy achieved through hybrid model contributes to a more reliable and effective forecasting model.

Integrating sentiment analysis enriches the hybrid model by considering public opinion from Twitter. This fusion of market sentiment with quantitative predictions enhances the model's understanding of stock behavior. By exploring how sentiments align with stock movements, valuable insights emerge. This integration captures not only numerical trends but also the emotional context of the market, providing a more holistic approach to stock price prediction. It allows the model to respond to both factual market data and the subjective aspects of investor sentiment for a more subtle analysis.

## 4. EXPERIMENTAL EVALUATION

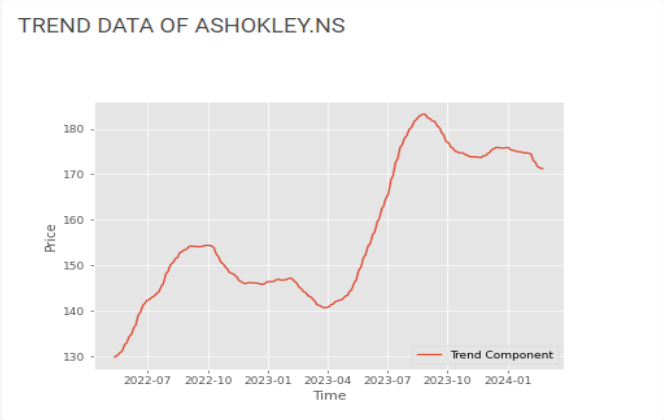
### 4.1 Environment Specification

The study has been done using x86-based processor. Furthermore, the machine has 8GB of RAM and 1TB Hard disk. The models are constructed using hybrid LSTM and ARIMA models in machine learning with the help of Keras and TensorFlow libraries available in Python.

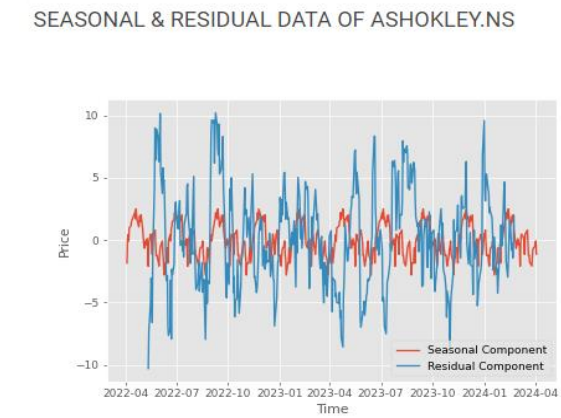
### 4.2 Dataset

The Stock data is decomposed into trend, seasonal and residual data using stats model library. The LSTM model is applied to both seasonal and residual data and ARIMA model is applied to trend data. Each of the data decomposed into trend, seasonal and residual component are trained individually. The decomposition of time series data is shown in Chart 1. The stock prediction is calculated using the formula:

$$Prediction = Seasonal\ prediction + Trend\ prediction + Residual\ Prediction \quad \text{.... (1)}$$




(a)




(b)

**Chart – 1:** Time Series Decomposition of Data
(a) Trend decomposition Graph
(b) Seasonal & Residual Component Graph

**Table -1:** Experimental Setup

| Process Name | S. No | Action |
|---|---|---|
| Input | 1. | Collect historical data of stock prices from Yahoo finance |
| | 2. | Collect tweets from Twitter for sentiment analysis |
| Environment Configuration | 3. | VS code |
| | 4. | Import all necessary libraries and packages |
| Data Preprocessing | 5. | It involves applying normalization, data reduction, data transformation, and data cleaning on the collected data |
| Training and Testing | 6. | Build a prediction model trained on hybrid LSTM and ARIMA |
| | 7. | The dataset is split into 80% for training and 20% for testing. |
| Model Compilation | 8. | Set 1000 epochs for Hybrid LSTM and ARIMA model training (0, 1, 0) |
| Performance Report | 9. | Generate prediction report |
| | 10. | Generate model accuracy with an error rate |
| Prediction | 11. | Predict tomorrow's stock price for a particular organization |

### 4.3 Analysis

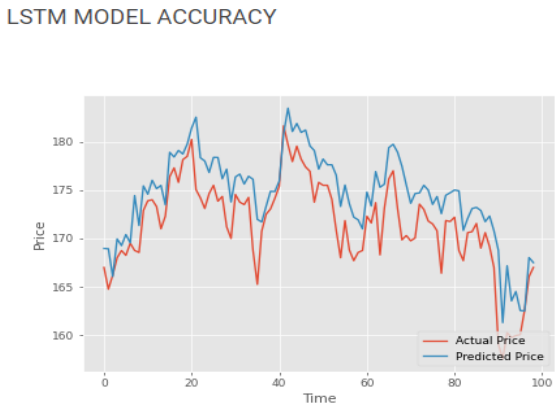
The hybrid LSTM-ARIMA model was trained and evaluated on the dataset using LSTM architecture with three layers, each comprising 50 units, and a dropout rate of 0.1. The model was optimized using the Adam optimizer with a mean squared error loss function over 1000 epochs. The AIC value is evaluated to obtain the better ARIMA model and configured in the order of (p, d, q) – (0, 1, 0).The performance of the hybrid model was assessed using several metrics. The following values represent the forecasted stock prices for Ashok Leyland. The root mean squared error (RMSE) for the hybrid model was 1.16. The mean absolute error (MAE) for the hybrid model was 0.58. Additionally, the mean absolute percentage error (MAPE) for the hybrid model was 0.01.

The accuracy of the hybrid model was evaluated based on its ability to predict the trend components of the time series data.

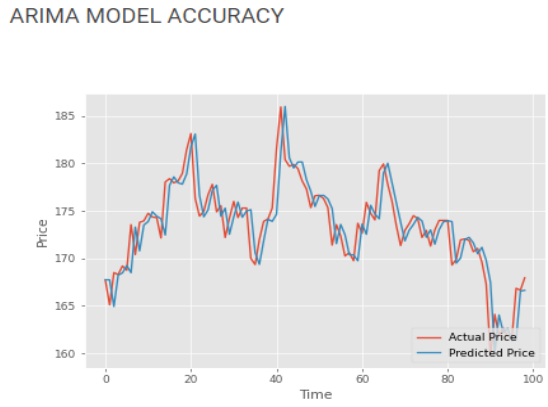
### 4.4 Deployment

Deployment of a stock market price prediction project involved setting up a user-friendly interface through a web-based platform built on the Flask framework. Users could access the application through a web browser to fetch the predicted price for a particular stock market. The data was fetched from Yahoo Finance, and preprocessing of data took place at the backend of the project. The flask backend handled the training and testing of data using a hybrid LSTM and ARIMA model. Sentiment analysis using the Natural Language Processing (NLP) technique was applied to process Twitter data and categorize the tweets into positive, negative, and neutral sentiments. Evaluation metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute

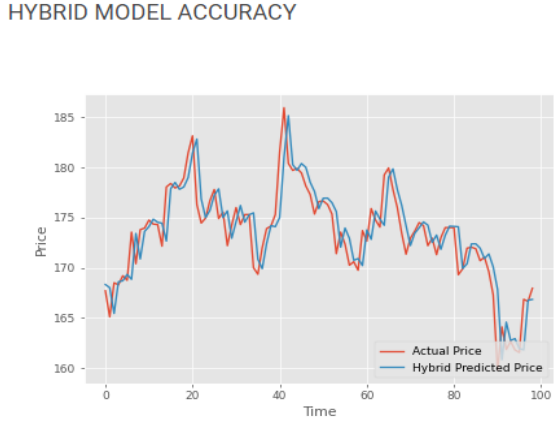
Percentage Error (MAPE) were utilized to assess the accuracy and performance of the stock price predictions generated by the Hybrid LSTM & ARIMA model. Table 2 specifies the final prediction result for Ashok Leyland limited. Utilizing Flask enabled seamless interaction between users and the application, providing a predicted value of a stock market price promptly and effectively. The Chart 2 illustrates the accuracy of each model in predicting the stock price of Ashok Leyland Limited.



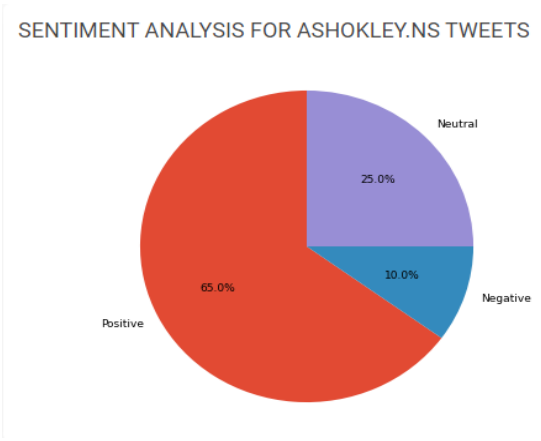

(a)



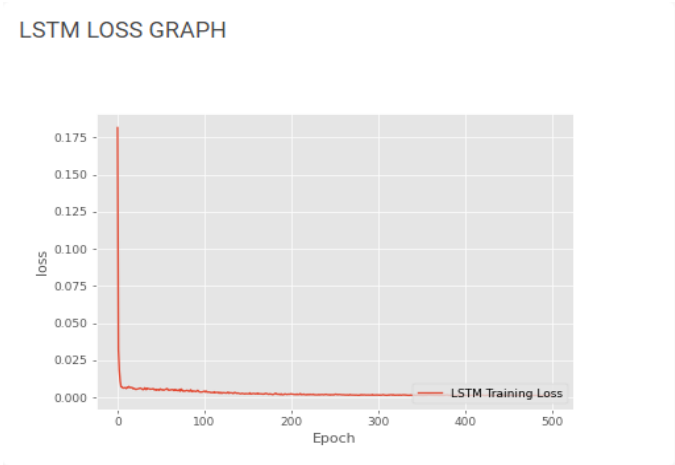

(b)



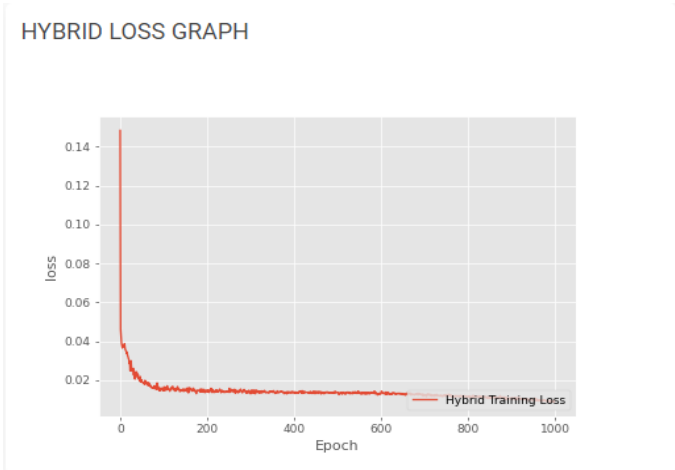

(c)




(d)




(e)




(f)

**Chart – 2:** Accuracy predictions of Different Model
(a) LSTM Model Accuracy for Ashok Leyland
(b) ARIMA Model Accuracy for Ashok Leyland
(c) Hybrid Model Accuracy for Ashok Leyland
(d) Sentiment Analysis chart for Ashok Leyland
(e) Loss graph for LSTM
(f) Loss graph for Hybrid model

**Table – 2:** Final prediction result of Ashok Leyland Limited

| Model | RMSE | MAE | MAPE |
|---|---|---|---|
| ARIMA | 2.48 | 1.83 | 0.03 |
| LSTM | 2.63 | 1.98 | 0.02 |
| Hybrid LSTM & ARIMA Model | **1.16** | **0.58** | **0.01** |

## 5. CONCLUSION

The paper aimed to predict stock market prices by combining forecasting techniques with sentiment analysis. The combination of Long Short-Term Memory (LSTM) and Auto-Regressive Integrated Moving Average (ARIMA) in a hybrid model is effective, demonstrating better predictions than using either model alone. The results indicated that the hybrid model was successful in reducing errors in predicting stock prices. The decomposition of data into trend, seasonal and residual and applying the data to the right component improves the accuracy of the hybrid model. Looking ahead, further research was suggested to strengthen sentiment analysis, explore new data sources, and utilize more advanced machine learning models to adapt to the dynamic nature of markets. Ultimately, the work contributed valuable insights for investors, analysts, and researchers seeking a practical and accurate approach to understanding and predicting stock market movements.

## REFERENCES

[1] R. Behera, S. Das, S. Rath, S. Misra, and R. Damasevicius, "Comparative Study of Real Time Machine Learning Models for Stock Prediction through Streaming Data," JUCS - Journal of Universal Computer Science, vol. 26, no. 9, pp. 1128–1147, Sep. 2020 , doi:10.3897/jucs.2020.059.

[2] K. Chen, Y. Zhou, and F. Dai, "A LSTM-based method for stock returns prediction: A case study of China stock market," 2015 IEEE International Conference on Big Data (Big Data), Oct 2015, pp. 2823-2824.

[3] S. Chen and L. Ge, "Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction," Quantitative Finance, vol. 19, no. 9, pp. 1507–1515, Jul. 2019, doi: 10.1080/14697688.20191622287.

[4] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," European Journal of Operational Research, vol. 270, no. 2, pp. 654–669, Oct. 2018, doi: 10.1016/j.ejor.2017.11.054.

[5] H. He and S. Dai, "A prediction model for stock market based on the integration of independent component analysis and Multi-LSTM," Electronic Research Archive, vol. 30, no. 10, pp. 3855–3871, 2022, doi: 10.3934/era.2022196.

[6] P. Mehta, S. Pandya, and K. Kotecha, "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning," PeerJ Computer Science, vol. 7, p. e476, Apr. 2021, doi: 10.7717/peerj-cs.476.

[7] C. Zhao, P. Hu, X. Liu, X. Lan, and H. Zhang, "Stock Market Analysis Using Time Series Relational Models for Stock Price Prediction," Mathematics, vol. 11, no. 5, p. 1130, Feb. 2023, doi: 10.3390/math11051130.

[8] H. Widiputra, A. Mailangkay, and E. Gautama, "Multivariate CNN-LSTM Model for Multiple Parallel Financial Time-Series Prediction," Complexity, vol. 2021, pp. 1–14, Oct. 2021, doi: 10.1155/2021/9903518.

[9] B. Reddy and U. J.C, "Prediction of Stock Market using Stochastic Neural Networks," International Journal of Innovative Research in Computer Science & Technology, vol. 7, no. 5, pp. 128–138, Sep. 2019, doi: 10.21276/ijircst.2019.7.5.1.

[10] H. N. Bhandari, B. Rimal, N. R. Pokhrel, R. Rimal, K. R. Dahal, and R. K. C. Khatri, "Predicting stock market index using LSTM," Machine Learning with Applications, vol. 9, no. 100320, p. 100320, May 2022, doi: 10.1016/j.mlwa.2022.100320.

[11] T. Mankar, T. Hotchandani, M. Madhwani, A. Chidrawar, and C. S. Lifna, "Stock Market Prediction based on Social Sentiments using Machine Learning," 2018 International Conference on Smart City and Emerging Technology (ICSCET), Jan. 2018, doi: 10.1109/icscet.2018.8537242.

[12] S. Antad et al., "Stock Price Prediction Website Using Linear Regression - A Machine Learning Algorithm," ITM Web of Conferences, vol. 56, p. 05016, 2023, doi: 10.1051/itmconf/20235605016.

[13] A. Q. Md et al., "Novel optimization approach for stock price forecasting using multi-layered sequential LSTM," Applied Soft Computing, vol. 134, p. 109830, Feb. 2023, doi: 10.1016/j.asoc.2022.109830.

[14] H. Kaeley, Y. Qiao, and N. Bagherzadeh, "Support for Stock Trend Prediction Using Transformers and Sentiment Analysis," arXiv.org, May 17, 2023. https://arxiv.org/abs/2305.14368.

[15] T. H. H. Aldhyani and A. Alzahrani, "Framework for Predicting and Modeling Stock Market Prices Based on Deep Learning Algorithms," Electronics, vol. 11, no. 19, p. 3149, Sep. 2022, doi: 10.3390/electronics11193149.

[16] G. Alfonso and D. R. Ramirez, "A Nonlinear Technical Indicator Selection Approach for Stock Markets. Application to the Chinese Stock Market," Mathematics, vol. 8, no. 8, p. 1301, Aug. 2020, doi: 10.3390/math8081301.

[17] J. Cao, Z. Li, and J. Li, "Financial time series forecasting model based on CEEMDAN and LSTM," Physica A: Statistical Mechanics and its Applications, vol. 519, pp. 127–139, Apr. 2019, doi: 10.1016/j.physa.2018.11.061.

[18] F. Feng, X. He, X. Wang, C. Luo, Y. Liu, and T.-S. Chua, "Temporal Relational Ranking for Stock Prediction," ACM Transactions on Information Systems, vol. 37, no. 2, pp. 1–30, Mar. 2019, doi: 10.1145/3309547.

[19] P. Xia, Z. Li, W. Zhang, and B. Li, "Adaptive Long-Short Pattern Transformer for Stock Investment Selection," Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Jul. 2022, doi: 10.24963/ijcai.2022/551.

[20] C. Zhao, X. Liu, J. Zhou, Y. Cen, and X. Yao, "GCN-based stock relations analysis for stock market prediction," PeerJ, vol. 8, pp. e1057–e1057, Aug. 2022, doi: 10.7717/peerj-cs.1057.