Comparison Between ARIMA and LSTM models in Stock Price Forecast

Ting Lin

Department of mathematics and statistics, Xi'an Jiaotong University, Shaanxi, China

Corresponding author: tinglin365@gmail.com

Abstract:

The improvement of model forecasting performance for long- standing multivariate time series is the primary focus of academic research on time series prediction that encompasses a range of methodologies. This paper aims to examine the performance of Autoregressive Integrated Moving Average model (ARIMA) and Long Short-Term Memory (LSTM) models in predicting timeseries data. Through the process of historical data, the author evaluated the precision and efficiency of these two models in forecasting Apple Inc.'s closing stock prices. The previous articles suggest that while the ARIMA model performs well with linear and stable datasets, the LSTM model, due to its deep learning architecture, can capture more complex patterns and interactions, leading to more accurate predictions in nonlinear and volatile markets. The study also discusses the trade-offs between the two models in terms of data requirements, computational resources, and model interpretability. These two models are pivotal in time series forecasting, providing robust frameworks for predicting future data points by analyzing historical patterns, with ARIMA focusing on linear dependencies through autoregressive and moving average components, and LSTM leveraging deep learning to capture complex, non-linear relationships.

Keywords: Time series forecast, Stock price, LSTM, ARIMA.

1. Introduction

In finance, economics, and meteorology, time series forecasting is a crucial issue. It is often done using the Autoregressive Integrated Moving Average model (ARIMA) model, which is formally named the Autoregressive Integrated Moving Average Model. It was introduced by Box and Jenkins in the early 1970s [1], The establishment of dynamic interdependencies within the data is employed by this model to forecast subsequent data points while predicting and analyzing univariate timeseries information. Owing to the simplicity as well as effectiveness in coping with linear timeseries data, the traditional ARIMA model is widely accepted. However, Long Short-Term Memory (LSTM) model was created by Hochreiter [2] in 1997 due to the arrival of deep learning technologies. Owing to the capacity to handle complex non-linear patterns, the LSTM model has gained attention.

In the literature review, the author found that Liu, Tang and Cheng et al. (2020) [3] proposed a method combining ARIMA and LSTM to enhance forecasting accuracy. Siami-Namini, Tavakoli, and Namin (2011) [4] carried out empirical studies to compare the capability of these models on numerous datasets, emphasizing the benefits of LSTM for handling data with long-standing dependencies. Additionally, Sagheer and Kotb. (2019) [5] emphasized the potential of LSTM in dealing with nonlinear time series data, while The ARIMA model‘s limitations in practical applications were discussed in (2017) [6].

Through empirical analysis, this study seeks to evaluate the outcome differences between the ARIMA and LSTM models in predicting Apple Inc.'s stock prices. The author first introduced the theoretical foundations of both models, then describe in detail the processes of data preprocessing, model construction, and result analysis. Finally, the article discusses the implications of these findings for time series forecasting practices and propose some possible dimensions for future research.

2. Method and Theory

2.1 ARIMA Model

The primary objective of the ARIMA model is to utilize historical data in predicting the future condition. Ariyo, Adewumi and Ayo showed the basic theory in [7]. It is believed that the labeled values at the specific time are affected by the labeled values over a period of time in the past as well as contingencies over a period of time in the past. The ARIMA model aims to uncover the possible timeseries patterns hidden behind the data through autocorrelation and differencing, which are then utilized to get insight on future data. $ARIMA(p,d,q)$ is commonly used to represent it. The p-value is used in the first (AR) part. This suggests that there may be a causal relationship among the instant value and the values at the past ? p? time nodes. the q-value is used in the second (MA) part. This suggests that there could be a correlation between the instant value and the error at the past q? time nodes. and the d-value is used in the third (I) part. This is the sequence of the differencing needed for this model. The respective roles of these three components are: The first part is utilized to handle the autoregressive part of the time series, which examines the effect of past observations on the current values; The second part can be utilized to polish nonstationary timeseries function by removing the trend and seasonality. The third part handles the moving average portion of the timeseries, and it reflects the effects of previous forecast errors on the current values. The mathematical expressions for these two components are described below:

$$AR: X_t = c + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_p X_{t-p} + \xi_t \tag{1}$$

and

$$MA: X_t = \tau + \epsilon_t + \delta_1 \epsilon_{t-1} + \delta_2 \epsilon_{t-2} + \cdots + \delta_q \epsilon_{t-q} \tag{2}$$

If the difference is not currently considered (i.e., $d = 0$), then The ARIMA model is a direct amalgamation of the AR and MA models, and the formula for it can be expressed as:

$$\begin{aligned} X_t = c + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_p X_{t-p} + \\ \epsilon_t + \delta_1 \epsilon_{t-1} + \delta_2 \epsilon_{t-2} + \cdots + \delta_q \epsilon_{t-q} \end{aligned} \tag{3}$$

where X_t is the considered instant timeseries data. The relationship between the instant value and past p time nodes are depicted by the parameters β_1 to β_p. The parameters δ_1 to δ_q are utilized to describe the correlation between the current values and the errors during the past q time points. ϵ_t is the error term at the point in time; c is a constant term, which can be zero in this model. It is necessary to first compare the data if the timeseries being discussed is unsteady.

2.2 LSTM Model

LSTM networks are a typical mode of recurrent neural network (RNN) that is meant to retain information for extended periods, which makes them especially fit for timeseries data prediction tasks like getting deep insight on stock price. This method is able to cope with the vanishing gradient problem of traditional RNNs by creating a gating mechanism that controls the flow of information. By utilizing input, output, and forget gates, the network can learn when to store, forget, or use information, seizing both short-term and long-standing dependencies manifest among these data.

The cell state is the base of LSTMs, which acts like a conveyor belt and allows information to flow through the network without deterioration over time. A set of gates are used to control the state of this cell, which controls the flow of information. An LSTM unit is consisted of the input gate, forget gate, and output gate.

To start with, the forget gate is tasked with deciding which information to give up from the cell state. A sigmoid activation function is employed to generate values range from 0 to 1, with 0 being the absolute discard and 1 being the absolute preservation of the information. In addition, the first (input) gate is responsible for determining the storage of new information. An update gate is responsible for choosing which values to update, and a candidate module generates a series of new values that can be appended to the state. Then, the network can use element-wise multiplication and addition to add new information and discard old ones. The network can add new information while discarding old through element-wise multiplication and addition in this process.

The output gate is used to determine what information is output from the cell state. The cell state is determined by a sigmoid function and a ?*tanh* function to generate a string of values ranged from -1 to 1. The output gate and cell state are combined in the consequent output. The formal basis of the LSTM architecture is the use of differential equations, which are commonly used in scientific and engineering disciplines. BPTT is an expansion of the typical back-propagation algorithm to train LSTM networks to deal with sequential data and time dependencies within the network. Roondiwala, Patel, and Varma successfully predicted stock prices in [8]. LSTM networks are capable of capturing long-standing dependencies in data and are therefore widely used in a variety of fields such as natural language processing, speech recognition, and time series analysis. Various modifications and extensions have improved their performance and applicability to various tasks.

2.3 Brief Comparison Between These Two Models

In contrast, the ARIMA model is a statistical method that uses past values, differences, and moving averages to predict future data points. It consists of three main sections: an AR section, which captures the dependence of present values on past values; an I section, which ensures that the series is static; and a MA section, which utilizes past errors to make predictions.

When comparing LSTM and ARIMA for stock price prediction, several factors come into play. LSTMs can capture complex patterns and interactions within the data due to their deep learning architecture, potentially leading to more accurate predictions, especially in non-linear and volatile markets. However, they require larger datasets and more computational power. ARIMA models are simpler, require less data, and are easier to interpret, but may not capture the full complexity of stock price movements as effectively as LSTMs.

In article [9], an empirical study by Yamak, Yujian and Gadosey showed that ARIMA model is more valid than deep learning-based regression model. However, in [10], the authors showed that the LSTM model outperforms ARIMA in prediction precision as measured by metrics such as RMSE which is calculated by the average of squared error and MAPE which is the average of percentage error in absolute value. Inconsistencies in data availability, computational resources and model interpretability may affect the selection between LSTM and ARIMA.

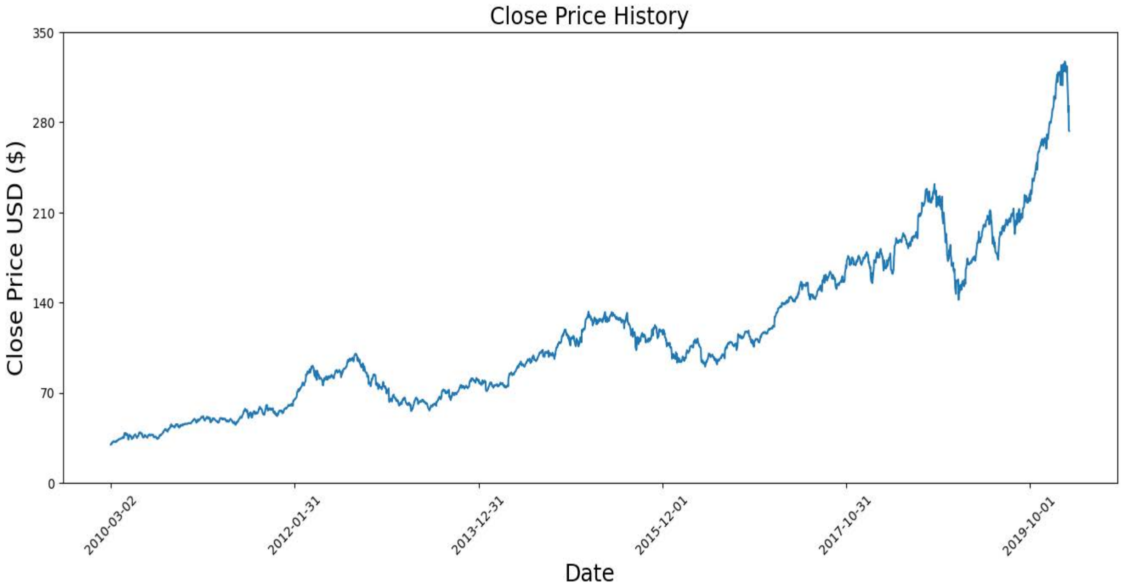

Fig. 1 The close prize of Apple Inc.'s stock from 2010 to 2020

3. Results and application

The authors initially approach the analysis from two perspectives: long-standing and short-standing forecasting. They employ two methodologies, namely the ARIMA

ISSN 2959-6157

model and the LSTM model, to examine the share price at close of stock market of Apple Inc. from January 1, 2010, to January 1, 2020 for the purpose of long-standing prediction. For the short-term forecasting, a subset of one year (2019-2020) is extracted to evaluate the performance of both methods. The database set used in this investigation is from Kaggle [11].

3.1 Data Preprocessing

The authors commence by loading the historical closing price data of Apple Inc.'s stock, subsequently employing matplotlib to plot the historical trend graph of the share price at close of stock market. Thereafter, the selected data are normalized using MinMaxScaler to facilitate subsequent neural network processing. The resulting data is depicted in the following two figures. The Fig. 1 shows the close prize of Apple Inc.'s stock from 2010 to 2020, while the Fig. 2 shows the close prize of Apple Inc.'s stock in year 2019.

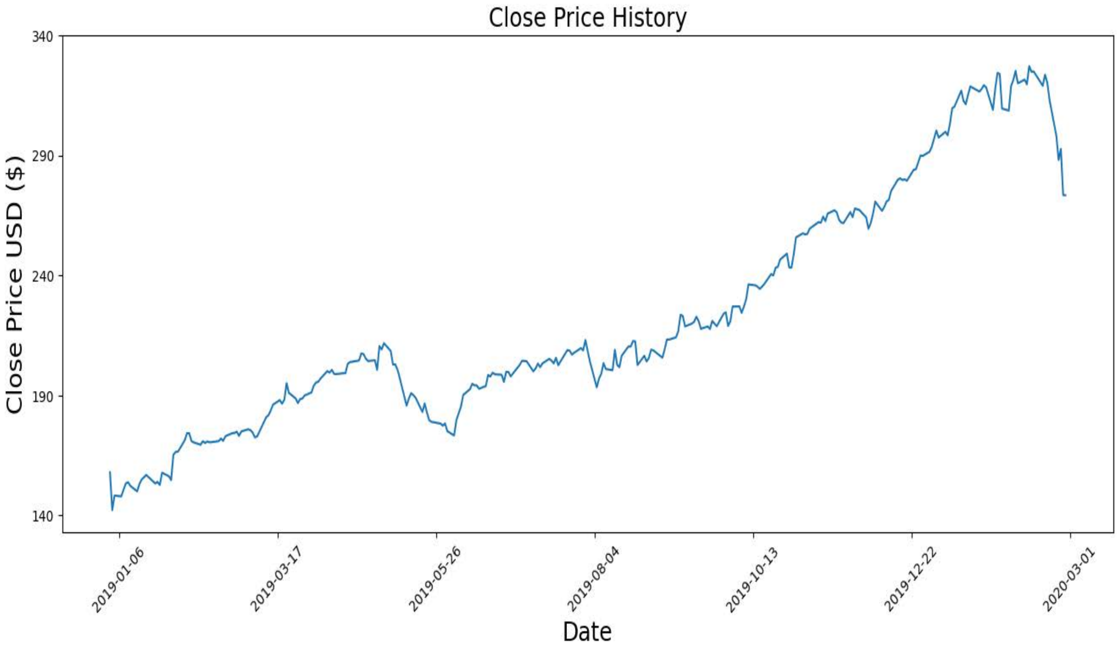

Fig. 2. The close prize of Apple Inc.'s stock in year 2019

3.2 Build the Model

Since examining whether the parameter d is zero is significant, the Augmented Dickey-Fuller (ADF) test is an important tool for determining the smoothness of the data in both the short-term and long-term situations.

The result of the above test shows that the ADF statistic is -3.326937, which is only greater than the critical value at 10% level of significance (-3.503, -2.893, and -2.584 at 1%, 5%, and 10% levels of significance, respectively). From this, it can be inferred that the data is smooth. In addition, the p-value is 0.013716, which is in line with common significance levels (e.g. 0.05 or 0.01). The authors can assume that the data is accurate because they have enough proof to reject the original hypothesis.

After autocorrelation and partial autocorrelation analysis, the arguments of the ARIMA model were determined by plotting the ACF and PACF plots. The coefficient plots for this particular dataset are clearly indicated in the Fig. 3. The best ARIMA model is selected using Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) by iterating over different permutations of p, d and q parameters. After training with the selected ARIMA model, the closing price is predicted for the next 30 days. The LSTM model requires the creation of both training and tested datasets, and then a neural network model consisting of two LSTM layers and a fully connected layer is constructed using Keras. The model is then trained using the mean square error as the damage function. The training model is used to make predictions on the test dataset and the RMSE is calculated between the forecasted and realized values.

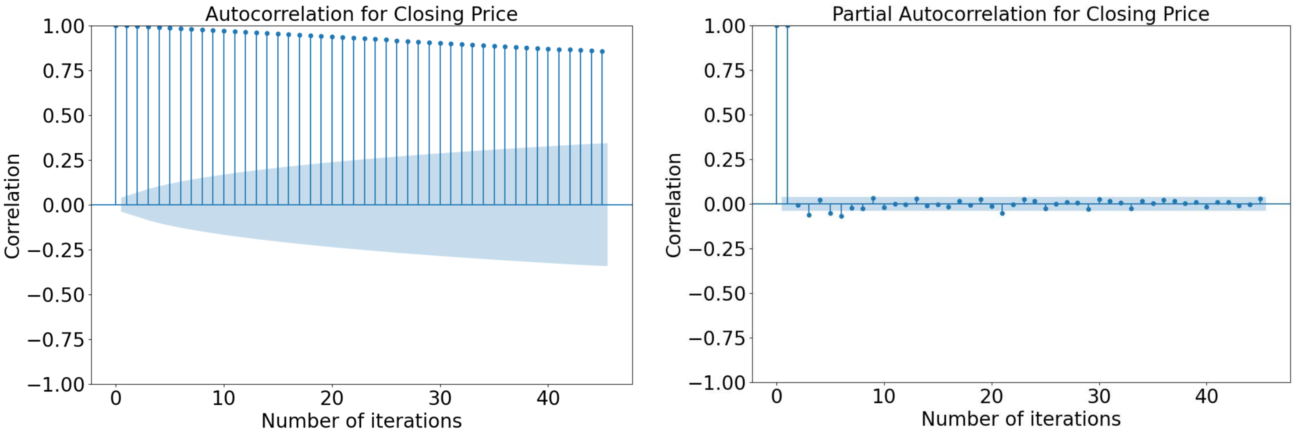

Fig. 3 The ACF and PACF graph of Apple Inc.'s stock dataset.

3.3 Relevant Results

The LSTM model's predictive outcomes are as follows: The best model in predicting on one-year data by method of ARIMA is *ARIMA*(3,0,3), its MSE is 196.86124858701308. when based on ten-year-long data, The best model is *ARIMA*(3,0,0), its MSE = 200.7489556944029. It can be known that the ARIMA model cannot efficaciously foresee fluctuations in data with poor stationarity, as it focuses on the weighted consideration of the conditions from a certain number of preceding days. Given a sufficiently extensive dataset, the LSTM model, due to its deep learning capabilities, can synthesize and analyze a vast array of data, thereby achieving superior accuracy. In contrast, this approach may not optimally utilize previous data for predictive purposes. Thus, it seems that the LSTM model has an advantage in predicting Apple Inc. stock prices. Results of prediction by LSTM model and ARIMA model are shown in Fig. 4 and Fig. 5.

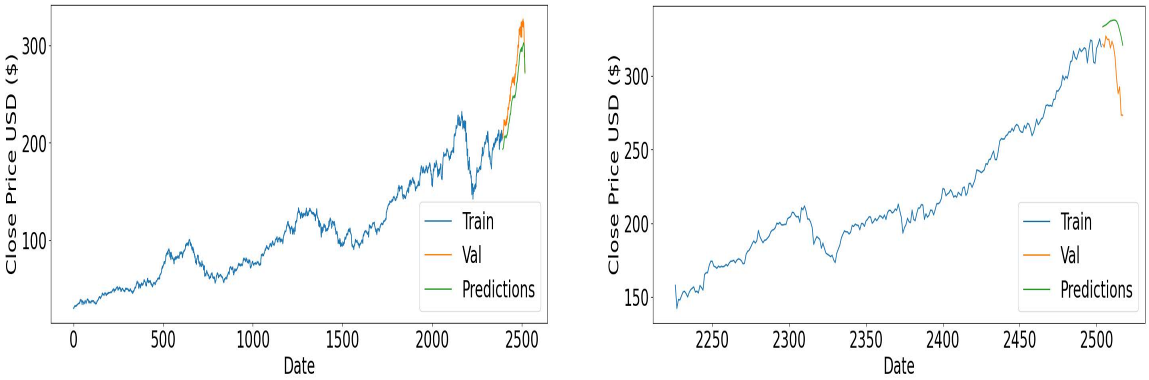

Fig. 4 Results of prediction by LSTM model

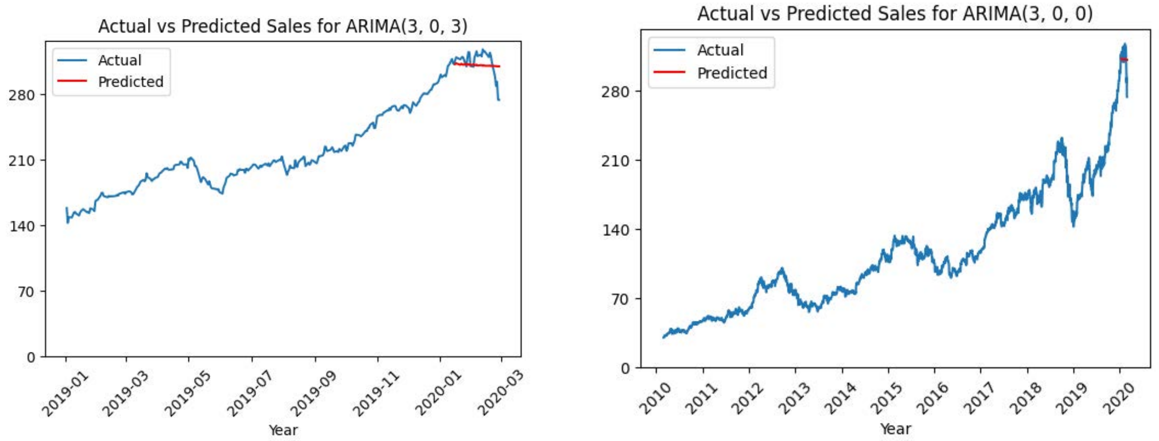

Fig. 5 Results of prediction by ARIMA model

3.4 Possible Reasons That Cause These Differences

The LSTM model uses the LSTM network, which is a deep learning model suitable for processing time series data of various data types. ARIMA model utilizes classical mathematical models for linear time series data and the ARIMA framework.

Data preprocessing and normalization have been carried out by both programs from a data processing perspective. However, the LSTM model typically requires a larger dataset for training, whereas the ARIMA model places greater emphasis on the stationarity and seasonality of the data. The LSTM model is more prone to needing more computational resources and time for training, while ARIMA model is more computationally elegant, thus the later can capture more complex non-linear patterns and may be better at predicting future prices, particularly when the data has non-linear characteristics. The ARIMA model performs well when the data are relatively stable and do not exhibit significant nonlinear features. Secondly, while the ARIMA model demonstrated robust performance with stable and linear datasets, it struggled to adapt to the dynamic nature of stock price fluctuations. This highlights the limitation of ARIMA in scenarios where the underlying data exhibits significant non-linearity and requires a more nuanced approach to prediction.

This study also highlights the critical importance of model choosing in timeseries forecasting. The nature of the information, the availability of computational resources, and the specific requisites of the forecasting assignment should be taken into account when choosing between ARIMA and LSTM. While ARIMA may be sufficient for simpler datasets, LSTM offers a more powerful alternative for datasets with higher complexity and non-linearity.

4. Conclusion

This study provides a comparison of ARIMA and LSTM models in timeseries forecasting, with a particular focus on the prediction of Apple Inc.'s stock prices. In conclusion, the findings suggest that LSTM models are the superior choice for time series forecasting in financial markets, where the data is frequently characterized by non-linearity and volatility. However, LSTM over ARIMA should be chosen after considering the advantages and disadvantages of computational resources, data requirements, and model interpretability. Discovering mixture models that combine the benefits of ARIMA and LSTM, as well as exploring their application in other domains beyond financial forecasting, could provide benefits for future research. It is important to mention that while the study provides valuable insights, it is not without limitations. The generalizability of these findings may be constrained by the specific dataset used in this analysis. Therefore, further research is needed to validate the conclusions across different datasets and market conditions. Deep learning and time series analysis's ongoing development has enabled the creation of more advanced models that could potentially surpass the enhancement of both ARIMA and LSTM in the future.

References

[1] Box, G. E. P., Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. Holden-Day.

[2] Hochreiter S. Long Short-term Memory [J]. Neural Computation MIT-Press, 1997.

[3] Liu B, Tang X, Cheng J, et al. Traffic flow combination forecasting method based on improved LSTM and ARIMA[J]. International Journal of Embedded Systems, 2020, 12(1): 22-30.

[4] Siami-Namini S, Tavakoli N, Namin A S. A comparison of ARIMA and LSTM in forecasting time series[C]//2018

17th IEEE international conference on machine learning and applications (ICMLA). Ieee, 2018: 1394-1401.

[5] Sagheer A, Kotb M. Time series forecasting of petroleum production using deep LSTM recurrent networks[J]. Neurocomputing, 2019, 323: 203-213.

[6] Stevenson S. A comparison of the forecasting ability of ARIMA models[J]. Journal of Property Investment & Finance, 2007, 25(3): 223-240.

[7] Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model[C]//2014 UKSim-AMSS 16th international conference on computer modelling and simulation. IEEE, 2014: 106-112.

[8] Roondiwala M, Patel H, Varma S. Predicting stock prices using LSTM[J]. International Journal of Science and Research (IJSR), 2017, 6(4): 1754-1756.

[9] Yamak P T, Yujian L, Gadosey P K. A comparison between arima, lstm, and gru for time series forecasting[C]//Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence. 2019: 49-55.

[10] Zhou K, Wang W Y, Hu T, et al. Comparison of time series forecasting based on statistical ARIMA model and LSTM with attention mechanism[C]//Journal of physics: conference series. IOP Publishing, 2020, 1631(1): 012141.

[11] Fares Sayah (2023), https://www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm/input.