Stock Market Analysis and Prediction Using LSTM

Yuhui Chen*

Department of Mathematics and Statistic, University of Toronto, Toronto, Canada

*Corresponding author: peachypie.chen@mail.utoronto.ca

Abstract. Even for professionals and analysts, predicting the value of stocks has proved to be a challenging endeavor. Because they shed light on the expected future path of the stock market, accurate prediction systems for the stock market are beneficial to traders, investors, and analysts. This is because traders, investors, and analysts can better anticipate the market's behavior. The increase in available choices for financial investments has contributed to the complexity and unpredictability of the stock market. The goal of this project is to develop a model that could precisely depicting the market's complexity as well as its high degree of volatility. The long short-term memory (LSTM) architecture of a neural network was implemented in this study to estimate Apple's next day closing price throughout the preceding decade. To forecast how the stock market will behave, its six fundamental indicators are integrated in a logical and well-balanced way. These indicators account for fundamental market data, macroeconomic data, and technical indications.

Keywords: Long short-term memory (LSTM), Prediction, Stock Prices, Recurrent neural network (RNN)

1. Introduction

The fluctuations in stock prices are difficult to forecast due to the presence of a number of interconnected elements. Despite this, the possibility of making a fortune on the stock market on a continuous basis increases as new technologies emerge. In addition to this, it assists specialists in identifying the signals that are the most important for producing more accurate forecasts. In order to maximise the return on stock option purchases while avoiding risk, having an accurate projection of the market value is essential. Following a thorough review of the industry, all players in the market aim to improve their revenues while simultaneously reducing their risks. The primary challenge consists in assembling all the information into a coherent whole and developing a robust model to provide accurate forecasts.

Predicting future stock prices is an undertaking that is difficult and complicated for firms, investors, and equity traders. The goal of this attempt is to forecast future earnings. The stock markets are notorious for their inherent chaos, non-linearity, noise, and lack of parametric predictability. Because of this, it is difficult to generate reliable and exact price projections for the future. The identification of relevant features from available financial data is an additional challenging aspect of stock prediction. There have been a few different solutions suggested for these issues. The information derived from a single time series was used in certain published works [1]. With the help of an appropriate model that is constructed using the appropriate characteristics, one can make fairly accurate stock price forecasts and obtain a better knowledge of the circumstances of the market. One of the models that is used most often for the processing of sequential input is known as a recurrent neural network, or RNN. The LSTM model, which is an upgraded neural network architecture for data of time series, is used to make predictions about the price of the stock index in this research project. In this article, statistics pertaining to the stock market will be investigated, with a specific focus placed on one technological stock, namely Apple's shares.

The subsequent sections of this article are organized as follows: The recommended strategy will be explored in Part 2. In part 3, the experiment to evaluate the effectiveness of the procedures will be conducted, and in part 4, conclusions will be reached.

2. Methodology

LSTM were initially described in, and since then, they have been successfully used in various applications that classify images and texts [2]. Now, the precise design of an LSTM layer is too extensive to detail without directing the reader to [3], which has the necessary information. Despite this, the key idea to keep in mind is that each layer is composed of LSTM units and that each of these units could regulate memory by means of an input gate, a forget gate, and an output gate.

The recurrent neural network improvement model known as LSTM [4]. Because they enable previously determined information to be utilized in current neural networks, they are said to have short-term memory. The most recent information is used for the activities that need to be done immediately. It's possible that one doesn't have a comprehensive collection of all the information that has come before about the neural node. LSTMs are a popular component in RNNs and other neural networks. Their effectiveness should be used for various sequence modeling difficulties across multiple application fields, such as natural language processing (NLP) and time series [5-8].

The computational framework of stock index price prediction is investigated via the data of time series augmented neural network LSTM model. Figure 1, and figure 2 show the research framework that has been suggested using a schematic that shows it from a bird's eye perspective. As seen in the picture, the model that will be developed via the proposed research will use primary, macroeconomic, and technological data that have been carefully chosen. Following that, the data that had been obtained were standardized using the technique of min-max normalization. After that, the input sequence for the LSTM model is created by using a certain time step in the process. The model uses hyperparameters such as the number of neurons, the period, the learning rate, the batch size, and the time step. The regularisation approach is used to ease the overfitting issue. Following the adjustment of the hyperparameters, the input data are then sent to the LSTM model, which then makes a forecast for the closing price of the stock market index. The RMSE and R square statistics were used to conduct the evaluation of the model's overall quality.

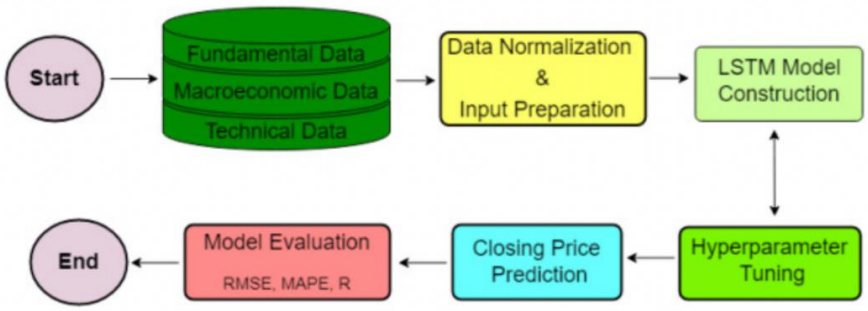

Figure 1. Flow Chart

The gradient vanishing issue is one of the most severe difficulties in RNN [9-11]. It is produced by RNN blocks utilizing the same parameters in each step, which leads to the problem. In order to find a solution to this issue, experiment will be conducted with different settings at each stage. In this particular instance, effort will be made to achieve a sense of equilibrium. Gradually introducing new parameters at each level while generalizing sequences of varying lengths and preserving the same number of learnable parameters. Then proposed gated RNN cells similar to LSTM cells.

Internal variables are referred to as gates, and a gating unit will always have them. The information available at every given time step, including those at the beginning of the process, is used to determine the value of each gate at that particular time step. The significance of the gates is then increased by multiplying it by a number of factors of interest that might affect them. A collection of values gathered at various time intervals throughout the course of a period; time series data enables us to monitor changes that have occurred over the course of a period. The data from time series may be used to monitor progress over more extended periods of time, such as milliseconds, days, or even years.

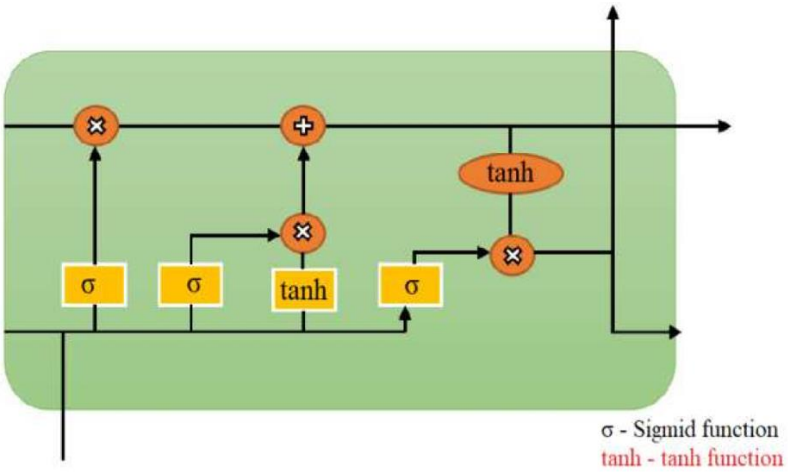

Figure 2. Long short-term memory (LSTM).

3. Experiment

3.1 Data

The data that were obtained from Kaggle and utilised in this experiment may be found there. The data set consists of 2518 records spanning from 2010-03-01 to 2020-02-28 and comprises six columns labelled High, Low, Open, Close, Volume, and Adj Close.

In order to do additional research, datatypes, missing values, and a great many other insights into the dataset and its properties will be retrieved. The values of the dollar sign that are missing will have a regular expression replaced for them, and the values will be typecast as float. After then, the data will be plotted so that the behaviour may be seen, shown in figure 3.

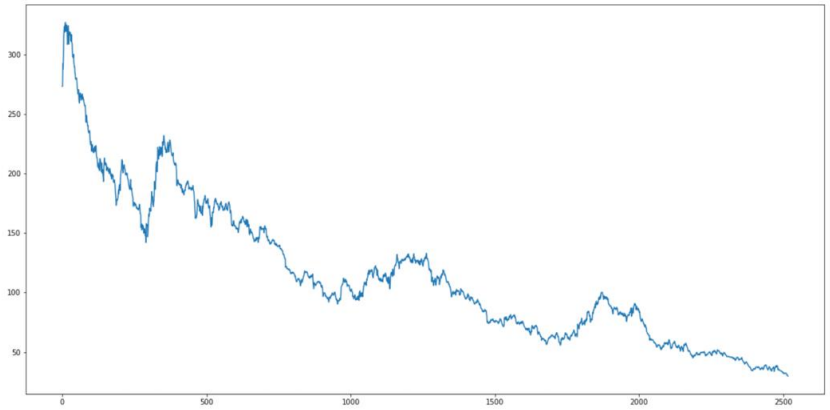

Figure 3. Visualization of missing value behavior

The data exhibits some seasonality as well as a discernible declining trend. However, in order to be certain about this seasonality, trend, and residuals, the stats model and the seasonal decomposition will be used in order to see the components of this time series, shown in figure 4.

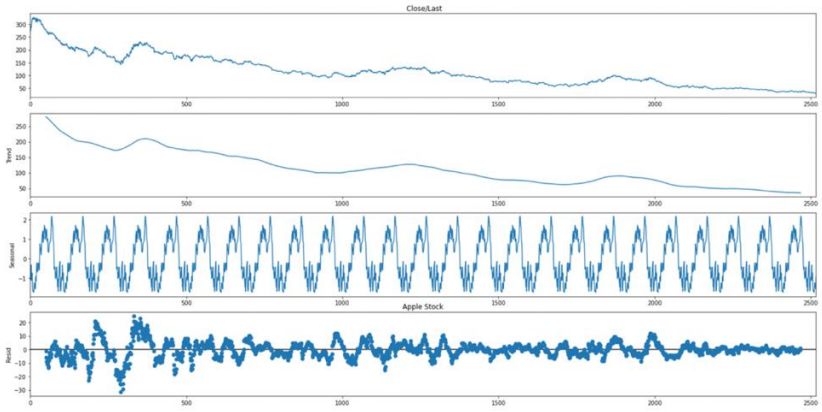

Figure 4. Visualization of the time series

3.2 Experiment Design

To demonstrate the efficacy of the LSTM model, the following experiments are conducted:
(1) First import reqired library, and then create dataframe
(2) Setting index, creating training and test sets
(3) Create and fit the LSTM model
MSE, R Square are selected as the index for evaluating the performance of the suggested approach; the index may be stated as follows:

$$MSE(F_{hat}) \coloneqq \frac{1}{n}\sum(yi - f\_hat(xi))^2 \tag{1}$$

$$R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2}$$

where RSS denotes residual sums of square and TSS denotes total sums of square
So, a general idea of the stock price data's behavior has been presented. Normalization is a very important part for any Recurrent Neural Network. For LSTM model, normalization will play an important role. Normalization using MinMaxScaler will bring the entire datapoints between a minimum and a maximum value. For this purpose, the values (0,1) will be used.
A 70%-30% train test split will be used. But before applying splitting, windowing the data is necessary. For that, store the training and testing data will be stored and their shapes in some variables. So, the training size is 1762 and test size is 756.

3.3 Experiment Results

The LSTM model may be fine-tuned for a variety of factors, such as modifying the number of LSTM layers, increasing the number of epochs, or adding a dropout value. But are the projections made by LSTM accurate enough to determine whether there will be a gain or a fall in the price of the stock? In order to construct the model, sequential, dense, and LSTM layers will be included. The model will be one that uses stacked LSTMs. That indicates there will be more than one LSTM layer in the network. The model will then be trained using the data once it has been fitted. In proportion to how well the model is trained, the accuracy of its predictions on the test data will be evaluated.
For the purpose of determining how effective the system is, the mean square error (MSE) will be used. The MSE value helps to bring the error, or the disparity, that exists between the goal and the actual output value down. The MSE is widely used and is an effective generic error measure for numerical prediction. Its use is ubiquitous. In comparison to other mean absolute errors, RMSE magnifies and severely penalizes mistakes that are already considerable. Determine the accuracy of the model by a mathematical examination of the error and the r2 score. The MSE came out to be 8.34, and the R-square was 0.93 after executing the algorithm. The model's predictions of the stock prices

are quite near to those basic values, as the mean squared error is just 8.34. A score of 0.93 for the R-squared metric indicates that the model prediction line is very well matched to the solid line.

The whole of the data, including both the training data and the test data, will be shown in figure 5:

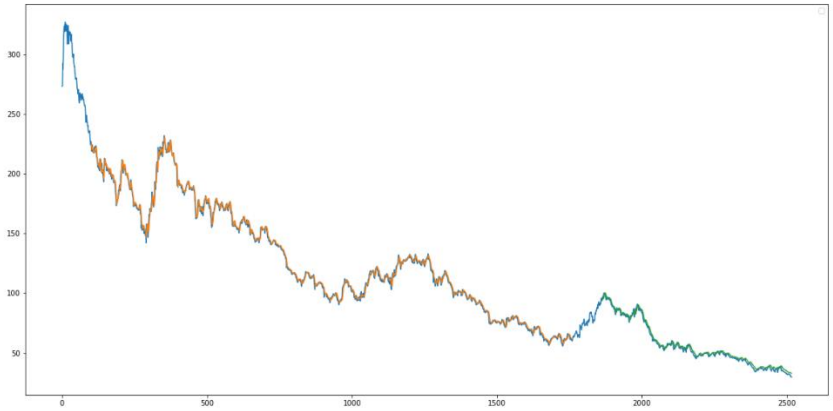

Figure 5. Model result

For predicting the future values, the values are stored and finally displayed and visualized in figure 6.

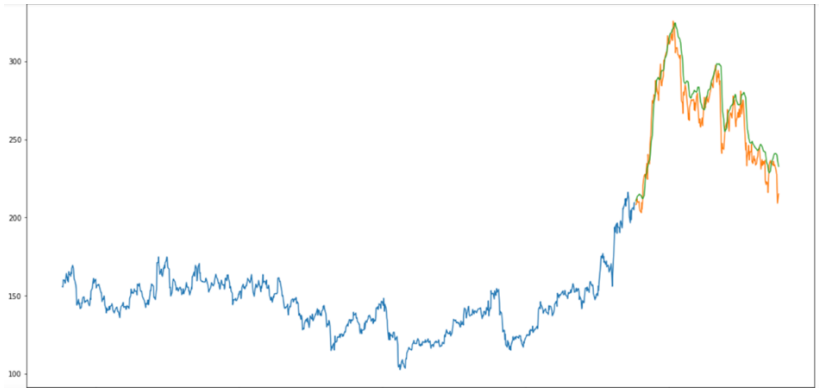

Figure 6. Predicting the future values

The model performed good at predicting the Apple Stock price using a Stacked LSTM model. A few alternative models will be evaluated based on the Apple stock data set. In this case, linear regression, SARIMA model, Prophet performed poorly compared to LSTM which its MSE is lower than the other models. Therefore, LSTM method made a better prediction. For reference, see in table 1.

Table 1. Comparison results

LMST MSE/R^2	Prophet MSE/R^2	Linear Regression MSE/R^2	ARIMA Model MSE/R^2
8.34	11.4	23.7	9.31
0.93	0.86	0.64	0.89

4. Conclusion

The art of accurately predicting future prices is gaining popularity among stock traders, individual investors, and portfolio managers. However, owing to the chaotic and nonlinear nature of the approaches, it is difficult to make stock price forecasts that are accurate and consistent. The projection may be affected by a variety of variables, including fundamental market data, macroeconomic data,

technical indications, and other considerations. Academics have been pushed to create new forecasting techniques by the increased popularity of trading on the stock market, and they have been adopting new methodologies. This approach of predicting is useful not just for academics but also for investors and everyone else who is interested in the stock market. To aid in the forecast of a stock index, a prediction model with a high level of accuracy is required. Using RNNs and LSTM units, which is one of the most accurate forecasting techniques, typically will be used it in this work. This method assists investors, analysts, and anybody else interested in investing in the stock market in predicting the future of the stock market.

References

[1] H. Jia, "Investigation into the efficacy of long short term memory networks for stock price forecasting," arXiv preprint arXiv:1603.07893, 2016.

[2] S.HochreiterandJ.Schmidhuber. Longshort-termmemory. Neural computation, 9:1735–80, 12 1997.

[3] C. Olah. http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[4] H. Goel, I. Melnyk, and A. Banerjee. arXiv preprint, 2017. R2N2: Residual recurrent neural networks for multivariate time series forecasting. Available at https://arxiv.org/abs/1709.03159".

[5] G. Petnehazi. arXiv preprint, 2019: Recurrent neural networks for time series forecasting. Available at https://arxiv.org/pdf/1901.00069.pdf.

[6] J. Roman and A. Jameel, "Backpropagation and recurrent neural nets in financial analysis of multiple stock market returns," in System Sciences, 1996., Proceedings of the Twenty-Ninth Hawaii International Conference on, vol. 2, IEEE, 1996, pp. 346-378.

[7] E. W. Saad, D. V. Prokhorov, and D. C. Wunsch, "Comparative study of stock trend prediction using time delay, recurrent, and probabilistic neural networks," IEEE Transactions on neural networks, vol. 9, no. 6, 1998, pp. 1456–1458.

[8] M.-C. Chan, C.-C. Wong, and C.-C. Lam, "Financial time series forecasting by neural network using conjugate gradient learning method and multiple linear regression weight initialization," Computing in Economics and Finance, vol. 61, 2000.

[9] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time series analysis: predicting and control. 2015, John Wiley & Sons

[10] Stock Price Trend Prediction Using Supervised Learning Methods by Sharvil Katariya and Saurabh Jain.

[11] Nelson et al., 2017, Stock market price movement prediction using LSTM neural networks, 2017 international joint conference on neural networks (IJCNN), IEEE (2017),