# Prediction of China Stock returns under the COVID-19 Pandemic based on the LSTM Model

Sijie Guo *

Donald Bren School of Information & Computer Sciences, University of California, Irvine

* Corresponding author. Email: sijieg@uci.edu

**Abstract.** The Coronavirus disease 2019, which is also known as the COVID-19, was first initiated in Wuhan, China, and then spread throughout the world, leading to an enormous quantity of reported cases and deaths. Accompanying the spread of the disease was the shrink in the economic markets. As one of the most important financial markets, the stock market was negatively impacted by the COVID-19, where many stock prices sharply decreased, and the performance of the stocks tended to be more unpredictable, compared to the performance of the stock market before the COVID-19 pandemic. China, as one of the most important economic entities, is also facing a decline in the stock returns in its stock market. This paper utilized the LSTM model to predict the China stock returns under the Coronavirus disease pandemic since 2019. The model is fitted by training on 12621 sequences, which were obtained during the Covid-19, and is tested with another 1260 sequences. The accuracy of the prediction in this paper is measured by MSE. The mean MSE after running the model fifty times is 0.0043.

**Keywords:** Coronavirus, COVID-19, China Stock Market, Machine Learning, LSTM, Stock Prediction.

## 1. Introduction

The stock market is a platform that assembles buyers and sellers to make exchanges among shares of public corporations, and it is influenced by an enormous number of factors including the political upheaval, interest rates, people's expectations, economic performance, and so on. It's a financial market that assembles trillions of dollars.

As the society embarked upon a COVID-19 pandemic era, the stock market throughout the world experienced a significant shrink and the performances of stocks became more unpredictable than usual [1]. Previous research analyzed data from the stock market from the top twenty worst-hit countries, which are selected based on the reported cases and deaths under the COVID-19 pandemic, indicating that the health news under the COVID-19 pandemic is leading to 'a negative and statistically significant effect on stock returns, indicating that returns decline as more information is sought on health issues since the pandemic outbreak [2].' Another previous research indicated 'that both the daily growth in total confirmed cases and in total cases of death caused by COVID-19 have significant negative effects on stock returns across all companies in China [3]. Furthermore, many other possible factors are causing the shrink of the stock return in China, including the reported cases and deaths, the shut-down of factories, the increase in the unemployment rate, the investors' pessimistic attitude, and so on. Considering the stock prices under the Covid-19 pandemic are influenced by many factors that were newly emerged, it is necessary that the training data, which is utilized to feed the prediction model in this paper, was collected entirely during the Covid-19 pandemic so that the prediction model can reach a higher accuracy in predicting the China stock returns under the Covid-19 pandemic specifically.

In a paper titled 'A LSTM-based method for stock returns prediction: A case study of China stock market', the author transformed the Chinese stock market's data into 30-days-long sequences accompanied by ten learning features and utilized LSTM to model and predict China stock return. The LSTM model in this paper increased the accuracy of prediction from 14.3% to 27.2%, compared to the model based on the random prediction method [4]. This paper implemented the 7767102 daily data of all 3049 stocks in the Chinese Stock Market from 1990/12/19 to 2015/09/10.

In this paper, 12621 sequences of daily records of the stocks, starting from the date 11th March 2020, which was the date when the World Health Organization (WHO) officially declared the Covid-19 as a pandemic [5], were fitted into the LSTM model. The purpose of this paper is to model and predict the Chinese stock returns under the COVID-19 pandemic based on the LSTM model. The accuracy of the prediction model is measured by Mean Squared Error (MSE).

## 2. Method

### 2.1. Data Description

The dataset in this paper is the historical stock data of the stock exchange in China stock market and is achieved from https://data.eastmoney.com/. The date range of this data is from 11th March 2020 to 11th March 2022. From this historical stock data, 12621 sequences were fed into the model for training, and another 1260 sequences were employed for testing.

### 2.2. Feature Extraction

The features of the dataset in this paper include the highest prices of the stock in one day, the lowest price of the stock in one day, the starting price of the stock in one day, the ending price of the stock in one day, and the volumes of each stock in each day.

In the research 'Predicting Stock Prices Using LSTM' by Murtaza Roondiwala, Harshal Patel, and Shraddha Varma, the LSTM model was employed to predict the stock returns of NIFTY 50 based on the previous five-year data of the NIFTY 50, and they made a comparable result using different parameters. The parameters they utilized were: i) Open/ Close ii) High/ Low/ Close iii) High/ Low/ Open/ Close, and when the High/ Low/ Open/ Close become the parameter, their LSTM model obtained the highest accuracy [9]. Thus, High/ Low/ Open/ Close is regarded as the first four features of the sequences. In another research paper written by A. Ronald Gallant, Peter E. Rossi, and George Tauchen, the volume is proved to be an important signal which indicates the upcoming changes in the stock prices, and 'four empirical regularities are found: (i) positive correlation between conditional volatility and volume; (ii) large price movements are followed by high volume; (iii) conditioning on lagged volume substantially attenuates the "leverage" effect; and (iv) after conditioning on lagged volume, there is a positive risk-return relation [10]' On account of the importance of the volume, the volume of each stock is considered as one of the fifth features in the dataset. Therefore, the features of each sequence in this research are the highest prices of the stock in one day, the lowest price of the stock in one day, the starting price of the stock in one day, the ending price of the stock in one day, and the volumes of each stock in each day. Notice that the 'Volume' reflects the total number of shares that are traded in one stock within one day.

The following table is derived from the dataset which is fitted into the LSTM model.

**Table 1.** Sample data of the utilized dataset

| Date | Feature1: Open | Feature2: High | Feature3: Low | Feature4: Close | Feature5: Volume |
|---|---|---|---|---|---|
| 2022/3/11 | 30.57 | 30.57 | 28.63 | 28.78 | 70997200 |
| 2022/3/12 | 28.41 | 29.54 | 28.23 | 29.23 | 87498504 |
| 2022/3/13 | 29.03 | 29.39 | 28.73 | 29.26 | 48012112 |

The dataset in this paper records the performances of the stocks of many companies in China. There are 12621 sequences in total which are utilized for training the model, and 1260 sequences are utilized for testing the model.

## 2.3. LSTM

### 2.3.1. The Advantages of LSTM

Coronavirus disease 2019, which is a highly contagious disease also known as the COVID-19, severely impacted the Chinese economy, leading to a significant increase in the unemployment rate. Many companies and stores were forced to shut down, even though they were actively changing their marketing strategies [6]. In China, many places such as cinemas, restaurants, shopping malls, and so on, had much fewer consumers than before. Furthermore, the shutdown of the raw-material companies led to the stop-production of many huge factories, such as Tesla factories in Shanghai [7]. The negative impacts of the COVID-19 on the economy were reflected in the stock market in China, leading to a significant decline. The Chinese stock market became more unpredictable. Many people, including fund managers, were facing a shrink in the financial property.

Neural networks have been proven to be capable of '[predicting] market directions more accurately than other existing techniques [8].' LSTM is an architecture utilized to predict the performance of future data based on the previous sequential data obtained. In this paper, LSTM is employed to make predictions of the future performance of the stock market in China during the COVID-19 pandemic era based on the LSTM networks.

LSTM networks, also known as the Long short-term memory networks, are one type of recurrent neural network for making predictions in data that are sequence-dependent. It is an advanced type of RNN architecture.

The hidden layer of LSTM is regarded as a gated cell. In the LSTM cell, previous intervals can be stored by the cell state, which is also known as the long-term memory. The forget gate, which is also known as the remember factor, is implemented below the cell state to adjust the cell state. The information that the forget gate required the cell state to forget is multiplied by 0 in a position of the matrix, and the information is remembered by the cell state when the forget gate produces an output of 1. According to the Forget Gate Equation,

$$f_t = \sigma_g\left(w_f \cdot x_t + U_f \cdot h_{t-1}\right) \tag{1}$$

the previous cell state forgets information accordingly based on the forget gate, and new information is installed based on the input gates' output.

In the cell state, part of the information will be forgotten and will not be able to pass to the next cell. Several gates inside the cell state are utilized to restrict the information from passing through directly.

The input gate, which is also called the save vector, is a sigmoid function that adds memory to the cell state. The range of the sigmoid function is between 0 and 1.

$$i_t = \sigma_g(w_i \cdot x_t + U_i \cdot h_{t-1}) \tag{2}$$

The input modulation gate contains a tan activation function with a range between -1 and 1. Because the output of the tanh activation function can be zero or a negative number, it enables the cell state to forget the memory. When the output of the tanh activation function is between 0 and 1, it adds memory to the cell state.

The output gate is called the focus vector that determines the values out of all possible values that should pass toward the next hidden state.

$$h_t = o_t \circ \sigma c_t \tag{3}$$

LSTM is an architecture that has a higher complexity compared to the RNN. Hence, an enormous number of resources and time are required to train to model, making it enable to solve complex

problems. On account of the unpredictable nature of the Chinese stock market under the COVID-19 pandemic, the size of resources and time in this paper had to be significantly larger.

LSTM is an architecture utilized to predict future values based on the previous data that are sequence-dependent. It has many advantages compared to the other models which are also implemented in predicting the sequence-dependent data, such as RNN and GRU.

When comparing LSTM with RNN, the first advantage of LSTM networks is that LSTM can remember values with random intervals, even if the interval is large. Under the reinforcement of the LSTM, the model can predict a sequence after an interval of 5000 without forgetting the starting point. In comparison, the RNN can only make a prediction of a sequence after an interval of 5. Inside one RNN cell, there are two inputs, including the time t and the output delivered from the last hidden state. The RNN cell forgets all the information from the past instead of the hidden state. Second, the LSTM networks contain three sigmoid activation functions and one tanh layer, while the RNNs contain only one tanh neural net layer. Third, LSTM has more room for adjustments than the RNN has, leading to better results.

When comparing LSTM with GRU (Gated Recurrent Units), the LSTM can limit the information passing through the cell and determine the new information embedded in the cell, while GRU cannot, since GRU does not store the cell state.

### 2.3.2. Training the LSTM Model

12621 sequences in the dataset are fitted into the LSTM model. The LSTM model learns to predict China's stock returns under the Covid-19 pandemic based on five features of the data: open, high, low, close, and volume. 1260 sequences are utilized for testing the model. The Optimizer (RMSprop) is utilized in this model. The learning rate is 0.0001 and the weight decay is 0. The learning rate is adjusted using stepLR.

## 3. Results

Mean Squared Error (MSE) calculates the means of the squares of the errors, and it is often utilized as a tool to measure the accuracy of an estimator in statistics. This paper employed MSE to measure the accuracy of the prediction of the LSTM model.

$$MSE = \frac{1}{M}\sum_{m=1}^{M}(p_m - \widehat{p_m})^2 \quad (4)$$

The number of epochs in this model is 100, and the model is run fifty times to obtain a mean MSE for describing the accuracy level of the prediction model. When Epoch = 0, the average Loss of the model is 3.4389.

The following figure shows the comparison between the stock returns predicted by the LSTM model and the stock returns in reality when the Epoch = 1

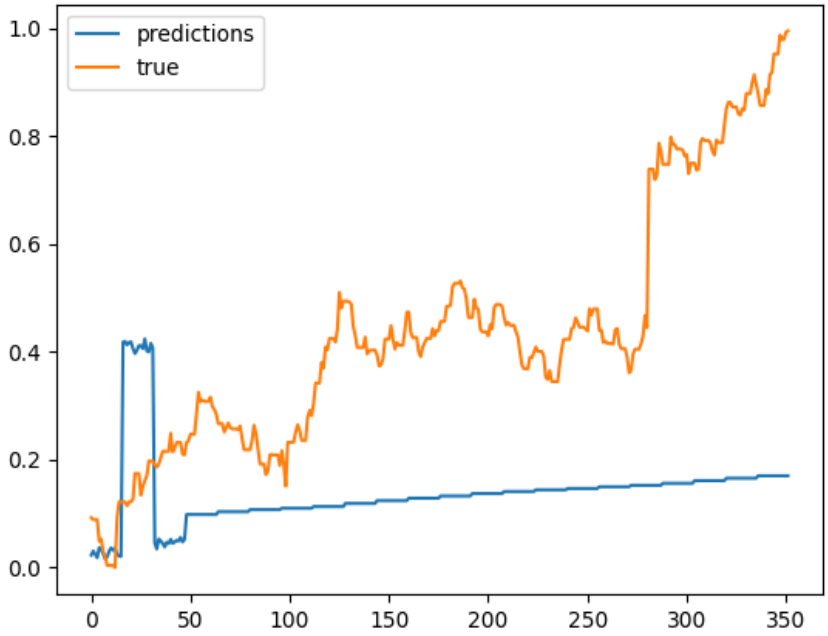


**Figure 1.** Evaluation results on Epoch1

The Y-axis represents the 'Date', and the X-axis represents the 'Stock Price'. After running for 100 epochs, the loss of the model decreased to 0.07487.

The change in Loss as the number of Epoch increases is shown below:

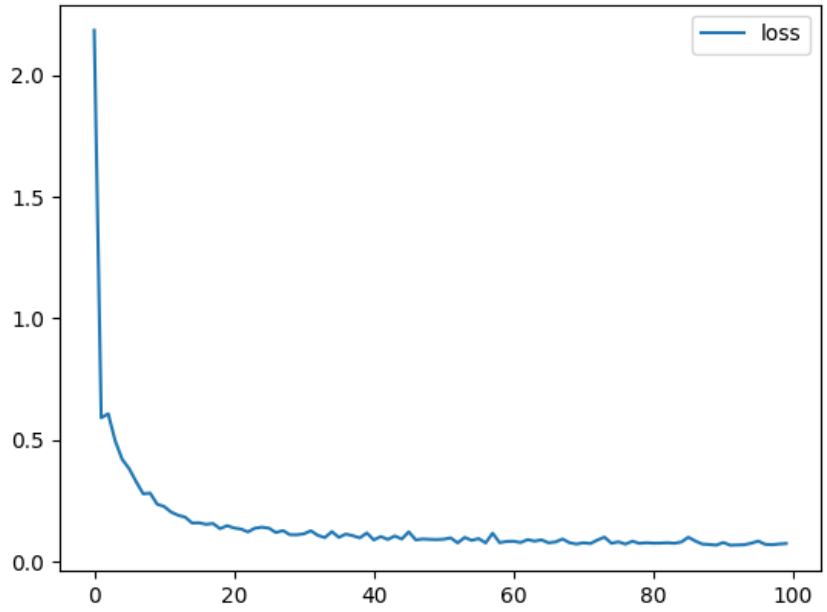


**Figure 2.** Learning Curve of LSTM network

The X-axis represents the number of epochs, and the Y-axis represents the loss. As is shown in the figure above, the model loss tends to decrease as the number of epochs increases, and the loss tends to be stabilized to a certain level as the number of epochs increases. On average, the loss of the model after 100 epochs is 0.07934.

The following figure exhibits one of the fifty eventual results of the stock returns predicted by the LSTM model, compared to the stock returns in reality.

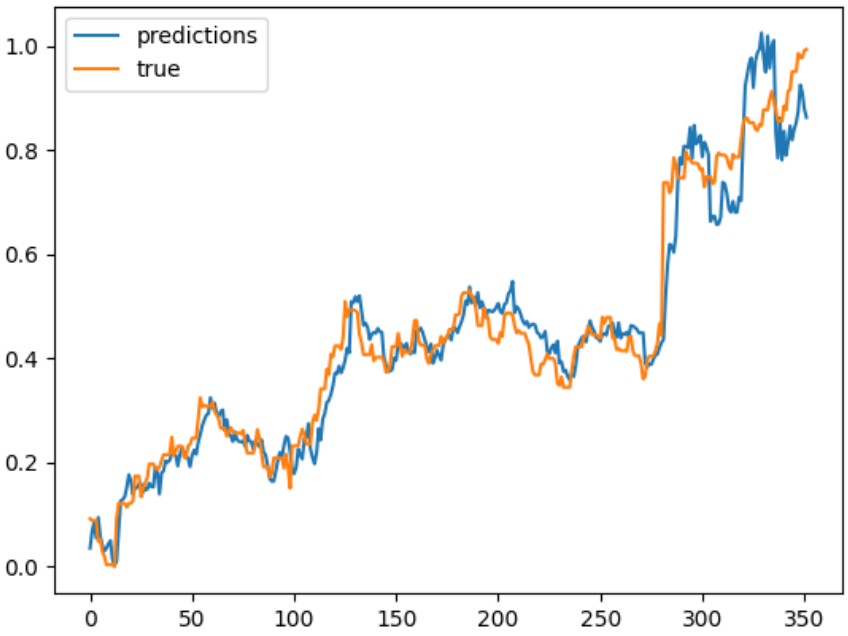


**Figure 3.** Evaluation Result on Epoch100

The blue line represents the predicted results while the orange line represents the true results. The Y-axis represents the 'Date', and the X-axis represents the 'Stock Price'. The MSE of this figure is 0.002953. The predicted result almost has the same tendency as the true stock performance.

Among the fifty times running the LSTM model, the lowest MSE was 0.002182, and the highest MSE obtained was 0.007317.

## 4. Conclusion

COVID-19 is a widespread disease which led to significant negative impacts on the economic markets throughout the world, as one of the most important economic entities, China also has to face a shrinking economy. The stock market, which serves as one of the most important financial markets experienced a decline, making a lot of investors lose money. In this paper, after fitting the LTSM model with 12621 sequences, the mean MSE this model obtains after running 50 times is 0.003753. The LSTM model in this paper enables stock investors and people who are pursuing careers that are related to the stock market to predict the performance of the China stock returns under the Covid-19 pandemic, even though investors cannot figure out which specific factors are causing the performance of the China stock returns to be unpredictable under the pandemic. The next step of this paper is to figure out the key factors that are causing the shrink of the stock market, which is a topic that may not relate to machine learning.

## References

[1] Verma, P., Dumka, A., Bhardwaj, A. *et al.* A Statistical Analysis of Impact of COVID19 on the Global Economy and Stock Index Returns. *SN COMPUT. SCI.* **2,** 27 (2021). https://doi.org/10.1007/s42979-020-00410-w.

[2] Salisu, Afees A., and Xuan Vinh Vo. "Predicting Stock Returns in the Presence of Covid-19 Pandemic: The Role of Health News." *International Review of Financial Analysis*, vol. 71, 2020, p. 101546., https://doi.org/10.1016/j.irfa.2020.101546.

[3] Al-Awadhi, Abdullah M., et al. "Death and Contagious Infectious Diseases: Impact of the COVID-19 Virus on Stock Market Returns." *Journal of Behavioral and Experimental Finance*, vol. 27, 2020, p. 100326., https://doi.org/10.1016/j.jbef.2020.100326.

[4] Chen, Kai, et al. "A LSTM-Based Method for Stock Returns Prediction: A Case Study of China Stock Market." *2015 IEEE International Conference on Big Data (Big Data)*, 2015, https://doi.org/10.1109/bigdata.2015.7364089.

[5] Cucinotta, Domenico, and Maurizio Vanelli. "WHO Declares COVID-19 a Pandemic." *National Library of Medicine*, 19 Mar. 2020, https://doi.org/10.23750/abm.v91i1.9397.

[6] Wang, Yonggui, et al. "Marketing Innovations during a Global Crisis: A Study of China Firms' Response to Covid-19." *Journal of Business Research*, vol. 116, 2020, pp. 214 – 220., https://doi.org/10.1016/j.jbusres.2020.05.029.

[7] Wen, W., et al. "Impacts of Covid-19 on the Electric Vehicle Industry: Evidence from China." *Renewable and Sustainable Energy Reviews*, vol. 144, 2021, p. 111024., https://doi.org/10.1016/j.rser.2021.111024.

[8] Yoo, P.D., et al. "Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation." *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, https://doi.org/10.1109/cimca.2005.1631572.

[9] Roondiwala, Murtaza, et al. "Predicting Stock Prices Using LSTM." *International Journal of Science and Research*, vol. 6, no. 4, Apr. 2017.

[10] Gallant, A. Ronald, et al. "Stock Prices and Volume." *Review of Financial Studies*, vol. 5, no. 2, 1992, pp. 199 – 242., https://doi.org/10.1093/rfs/5.2.199.