EE5183 FinTech Final Project Report – Stock Prediction feat. COVID-19 (Group 8)

Wei-Chen Hsu, Chi-Jung Huang, Ching-Hua Chou, Cheng-Wei Tsai, Chan Poh Yuen, Kai-Jie Zhang

ABSTRACT

Since January 2020, a new disease has caused an enormous catastrophe all over the world. It is named COVID-19. The outbreak of this pandemic caused over millions of deaths, and affected our life. Many countries announced lockdown, thousands of people lost their jobs, hospitals are flooded with patients. Global economy is also damaged. In the US, the stock market was devastated in spring. As the virus spread all over the world, the prediction of future economic performance is more difficult. In this project, we want to figure out the effect of infected cases to the stock market. We collected data of stock index and COVID-19 cases of several countries selected. In addition, we try to predict the stock price of these countries, with consideration of the effect of COVID-19. We propose two models, Bi-directional LSTM and DNN-LSTM, which are two popular models in stock price prediction. Then, we compare Taiwan stock market to other's and find out which kind of stock market is affected by the pandemic most. Moreover, we are curious about the proportion of historical stock index and COVID-19 statistics in the prediction of future stock price. We will evaluate that whether COVID-19 is suitable for stock price prediction.

Key Words—stock market, COVID-19, prediction, DNN, LSTM, Bi-directional LSTM

I. INTRODUCTION

Infectious diseases have always been a threat to humanity, especially those about which little or nothing is known. Since the last global pandemic, the Spanish flu, outbroke in 1918-1919, causing millions of deaths, people have tried to observe the impact of this disease to the economy. Although the data in early 20th century is rare, some situation has already been noticed, such as large amount of unemployment, many bankruptcies of corporations, enormous loss in several industries, etc. Pandemics like COVID-19 will surely have a significant influence on the global economy, as well as an impact on the financial markets. From 24 to 28 February 2020, stock markets worldwide reported their largest one-week declines since the 2008 financial crisis.

The stock market tends to be nonstationary, nonlinear, and highly noisy. Therefore, it is a very difficult task to forecast the stock index. Deep learning has been proven to predict the stock price successfully. Several approaches such as recurrent neural network (RNN) and long-short term memory (LSTM) have been applied in financial field. One of the algorithms that may be efficient in stock market forecasting is a bidirectional long short-term memory (BDLSTM). This algorithm is a combination of the bidirectional recurrent network (BDRNN) and long short-term memory (LSTM) cells. Another algorithm we consider is advanced deep neural network (DNN). Deep nonlinear topologies of DNNs can be built by stacking the hidden layers of the used nonlinear activation functions and used to analyze and model data with complex structures. In addition, we choose LSTM as nodes of this model, because these nodes are effective and expandable when used to address a number of problems involving series data.

First, we do the preprocessing of normalization on stock index and COVID dataset. Second, the dataset is grouped by choosing the size of training period. Third, these two models are used for training, with the target of the closed price. Last but not least, the model of each country with best performance are chosen for comparison. Advanced analysis is done to determine whether the model is good enough to predict future index.

II. MATERIAL AND METHOD

This section provides a detailed description of datasets used for forecasting price movements as well as a brief overview and mathematical description of bidirectional long short-term memory network and DNN-based prediction model.

A. Dataset Description and Model Evaluation

In order to observe various kinds of stock markets, we select 11 different countries: Taiwan, Japan, Singapore, UAE, Germany, UK, Italy, US, Brazil, South Africa, and Australia. These countries are located widely, and the stock markets are in different scale. In addition, the impact of COVID-19 is also different. Unlike Taiwan, the number of confirmed cases is less than a thousand, there are more than 10 million cases in the US. In order to compare the index between different countries, we select the most representative stock index for each country. The stock index we select is listed in Table 1. For stock market data,

Wei-Chen Hsu (R08942025), Institute of Communication Engineering, National Taiwan University.

we consider the following 9 features: high, low, open, close, volume, 10-days moving average, 30-days moving average, K, D. For COVID-19 data, we select the daily confirmed case as the input.

TABLE I
Selected Countries and Stock Index

Country	**Stock Index**
Taiwan	TAIEX
Japan	Nikkei225
Singapore	STI
UAE	ADI
Germany	DAX
Italy	FTSEMIB
UK	FTSE
US	NASDAQ
Brazil	IBOVESPA
South Africa	SOLJ
Australis	S/P

We choose the data mentioned above from February, 2020 to November, 2020. In this period, we are able to collect detail information of COVID-19. The data of first 8 months is used to train models, and the data of last 2 months is used for testing. Data normalization is implemented for preprocessing. All the training datasets are shuffled to prevent from overfitting. For the size of input data, we grouped data series of 2-30days. It is considered as a parameter in the training process. The criteria of a model is mean-square error (MSE).

B. Bi-directional Long Short-Term Memory

The main idea behind RNNs is to use sequential data as input. The RNN model can be simplified by unfolding the RNN architecture over the input sequence of data. However, conventional feedback neural networks process the data in one direction only, but in certain areas, past and future information is desirable. Therefore, bidirectional recurrent neural network (BRNN) is introduced. The basic idea was to extend the RNN architecture by introducing additional hidden layers where data were placed in the opposite, negative direction. Unfortunately, as a major drawback, BRNN in its basic form cannot model a complex time dynamic and it can suffer from the vanishing or exploding gradients.

One of the solutions to overcome the aforementioned problems is to use bidirectional long short-term memory (BDLSTM) architecture. One of the architectures of BDLSTM is shown in Figure 1. Such architecture differs from the RNN architecture in terms of hidden layers. BDLSTM has a LSTM cell as hidden layer, which consists of three gates: an input gate, a forget gate, and an output gate. The result of combining BRNN with LSTM cells is a BDLSTM network, which can model more complex time dynamics and deal with long-term dependencies. By using inputs in a positive sequence, the forward layer output sequence is calculated, and by using reversed inputs, the backward layer output sequence is calculated. Each element in the output vector of BDLSTM layer can be calculated by utilizing the σ function. In this project, BDLSTM is trained in order to predict price movement for the time period where the impact of COVID-19 on the global economy is relatively high. In the output vector of a BDLSTM layer, the last element is the predicted value for the next time iteration.

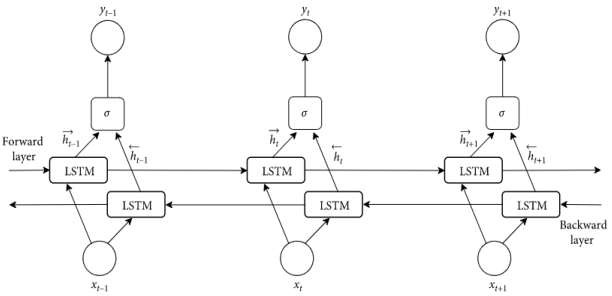


Fig. 1: The architecture of an unfolded BDLSTM

C. DNN-based Long Short-Term Memory Model

In an LSTM nerve cell model, the input gate determines how much information can be added to the nerve cell, and the output gate determines how much information can be output after processing. When used to model series data, LSTM NNs are able to capture mid- and long-term data and will not cause a time scale gradient vanishing problem, unlike earlier RNNs. Hidden layer architectures are mostly designed based on experience in previous studies. Hence, to reduce the number of comparison parameters in the experiment, all the hidden layers of the model are set to contain the same number of nodes. One of the architectures of DNN model is shown in Figure 1.

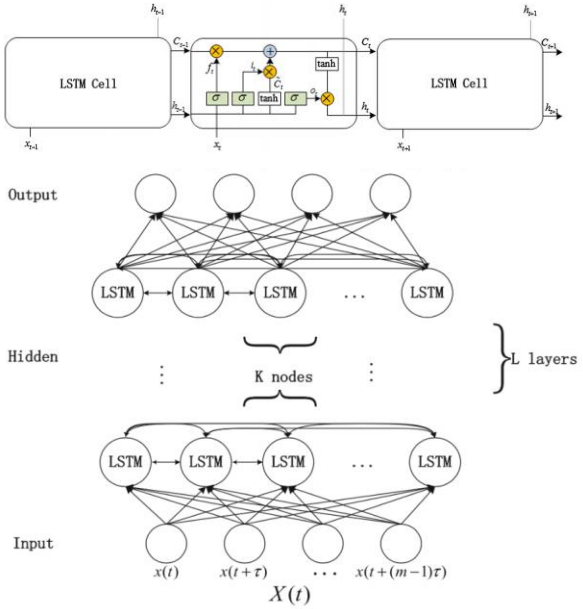


Fig. 2: Structural diagram of the DNN-based LSTM price prediction model

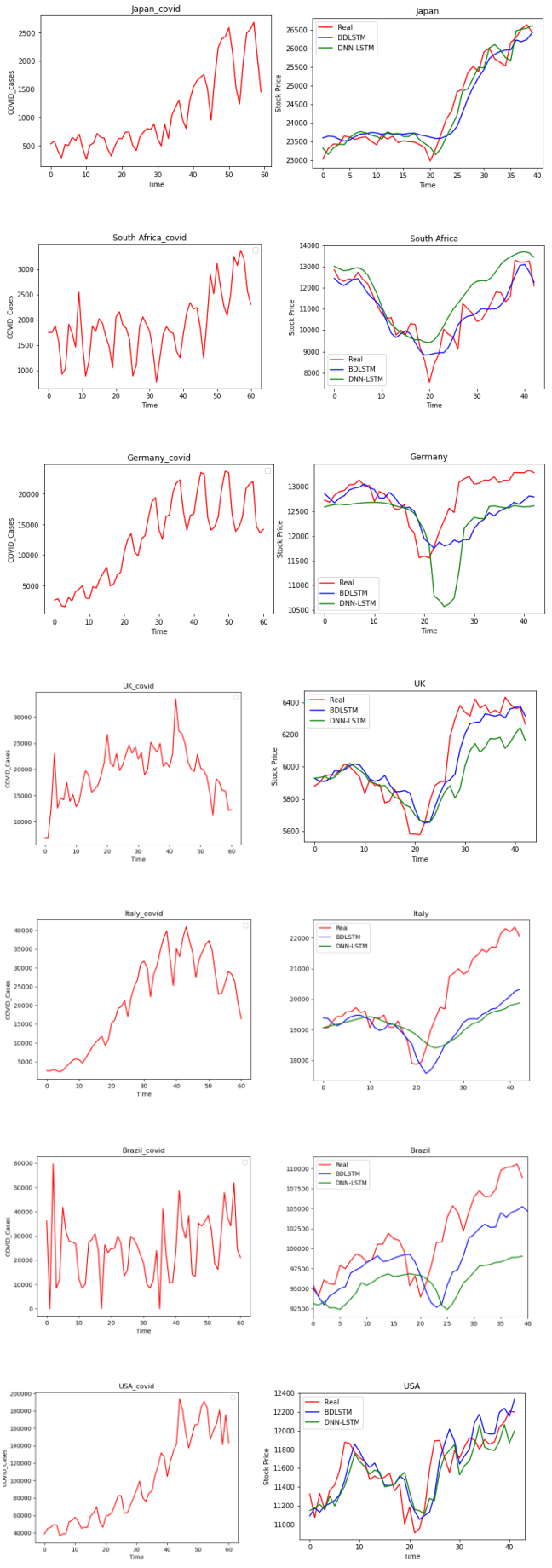
III. Evaluation

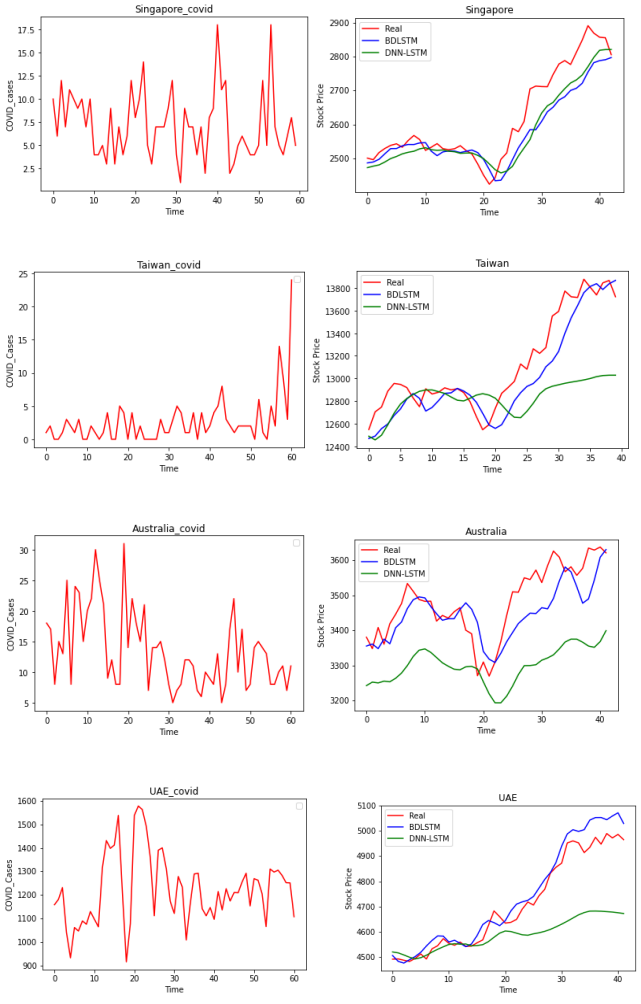
For each country, two models are applied for training and predicting respectively. The result is shown in Table 2. The comparison of real stock price with predicted one is shown in Figure 3.

TABLE II

Simulation results and performance comparison of different countries. Two models are used.

Country	**MSE(BDLSTM)**	**MSE(DNN)**
Taiwan	34558	200340
Japan	93024	69380
Singapore	4142	3672
UAE	2245	28375
Germany	596048	300909
Italy	1512497	3470509
UK	13247	33443
US	45948	48837
Brazil	4475089	3475292
South Africa	247961	986201
Australia	4924	37638





Generally, BDLSTM model has better performance. However, there are still exceptions, such as Germany, Japan, and Brazil. The result shows that the backward series is helpful for the prediction. In addition, we can find out that countries with more confirmed cases are more difficult to predict, such as Germany, Italy, Brazil, which has 10000 or more cases a day in October and November.

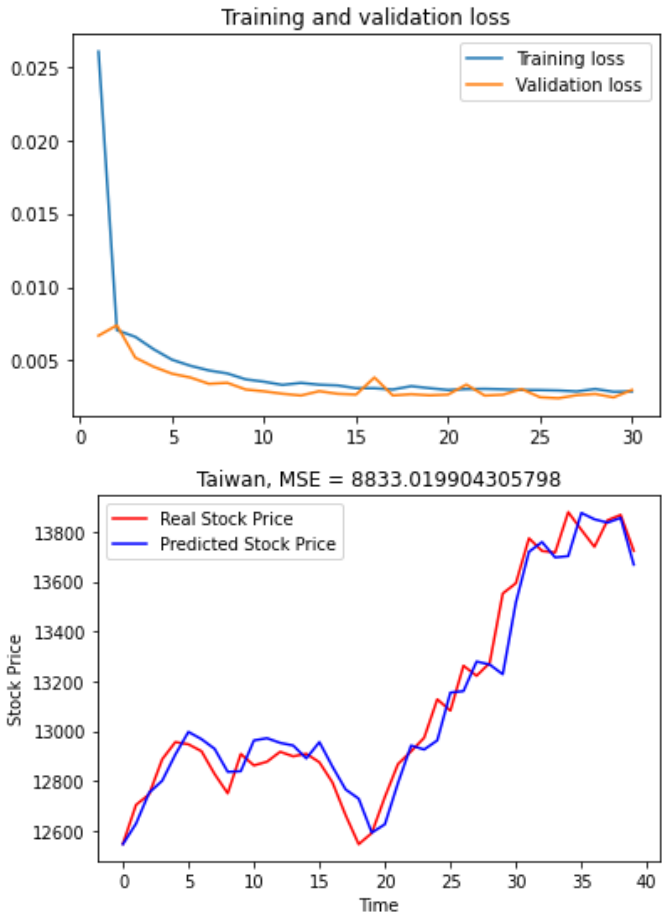
IV. Discussion

By observing the stock index data and COVID-related statistics, it is obvious that COVID-19 dataset has less and less effect on the stock market since the enormous declines in March. One possible reason is: all the governments announced several policies to protect the stock market. For example, the Fed announced unlimited quantitative easing (QE) in March. This operation affects the open market a lot. In addition, the industry structure of a country is also an important factor. For example, the demand in computers and communication applications rose rapidly due to working from home policies. It has made the stock price of companies in semiconductor and relative corporation skyrocket.

The MSE of each prediction is not good enough, due to lack of training data. For instance, there are less than 300 datasets for training for each stock market, and less dataset is available if we choose a larger window of time series. One of the solutions is expanding dataset. The stock data of longer period is considered, while the number of COVID-19 confirmed cases is defined as 0. Therefore, the expanded dataset is utilized for training, from January, 2003. The number of training dataset becomes 3000, much more than the previous one. The features selected for training and testing are also adjusted. The original selected stock-related features are high, low, open, close, volume, 10-days moving average, 30-days moving average, K, D. The feature “volume” is removed, due to the lack of relationship to the close price.

The revised training data is utilized for training BDLSTM model, with the determination of the following parameters: the number of neurons, the window size of time series, the activation function, the number of epochs and batch size, etc. In this research, the dataset of Taiwan is considered to examines the improvement of this model.

The result is shown in Figure 4. It is obvious that the MSE of the prediction declines. With more dataset used, the prediction is more stable, and the model performs better with less valid loss. However, the effect of COVID-19 is less than the previous model. There is just a small proportion of data with the confirmed COVID-19 cases larger than 0. The result may be the same as model without COVID-19 features, or even worse. The prediction of stock price in 2020 may also be worse, due to this pandemic is considered just a little bit. It is not the target of this project. Therefore, it is necessary to find an approach, which contains larger dataset and keeps the feature of COVID-19.



References

[1] D. Štifanić, J. Musulin, A. Miočević, S. Baressi Šegota, R. Šubić, and Z. Car, "Impact of COVID-19 on Forecasting Stock Prices: An Integration of Stationary Wavelet Transform and Bidirectional Long Short-Term Memory," Complexity, vol. 2020, p. 1846926, 2020/07/20 2020, doi: 10.1155/2020/1846926

[2] Yu, Pengfei & Yan, Xuesong. (2020). Stock price prediction based on deep neural networks. Neural Computing and Applications. 32. 10.1007/s00521-019-04212-x.