A STUDY IN SP500 COMPANIES STOCK CHANGES

Hung Vo
Faculty of Computer Science, Ho Chi Minh University of Technology

Abstract: The study aims to accurately predict the changing trend of the stock price of SP500 companies. In building a financial forecasting model, historical data and learned parameters of past years are used to predict future stock prices. Because of the complex datasets of the real world, machine learning models are no longer suitable for stock financial forecasting models. LSTM prediction models are used to predict the stock change trend on company data training. Through the forecast trend analysis under different models, LSTM predicts that the stock change trend of the enterprise model is closest to the changing trend of the actual earnings price. The prediction accuracy is better than other prediction models. Furthermore, this study also explores the relation between the stock change trend among companies. In general, stock correlation refers to how stocks move in relation to one another. While we can speak generally about asset classes being positively or negatively correlated, we can also specifically quantify correlation.
Keywords: Stock change price forecasting, LSTM, relation

1) Introduction:

In recent years, scholars have been exploring the stock market. Stock market is an important part of national economic development. In recent years, with the extensive application of deep learning technology, many domestic and foreign researchers began to use deep learning to conduct stock prediction research. The LSTM model is a special type of structure of the RNN model, in which three control units of the forgetting gate, the input gate and the output gate are added. As the information enters the model, the control unit in the model will make judgments on the information, leaving the conforming information and discarding the non-conforming information. Based on this principle, LSTM can solve the problem of long sequence dependence in neural networks.'

2) LSTM Stock Price Time Series Prediction Model:

2.1) Data Pre-Processing:

In the research of stock price, there are many real factors that affect the stock price, this results in the unstable data. Therefore, Standardization (Z-Score Normalization) is applied to the data in order to get a more stable dataset. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. Furthermore, applying this feature scaling in the preprocessing step can help the training step of the LSTM model converge faster.

2.2) Stock Price Forecasting:

Most data of the stock market is time series data, and the LSTM neural network has obvious advantages in processing time series information. The central role of an LSTM model is held by a memory cell known as a 'cell state' that maintains its state over time. The cell state is the horizontal line that runs through the top of the below diagram. It can be visualized as a conveyor belt through which information just flows, unchanged. Information can be added to or removed from the cell state in LSTM and is regulated by gates. These gates optionally let the information flow in and out of the cell. It contains a pointwise multiplication operation and a sigmoid neural net layer that assist the mechanism.

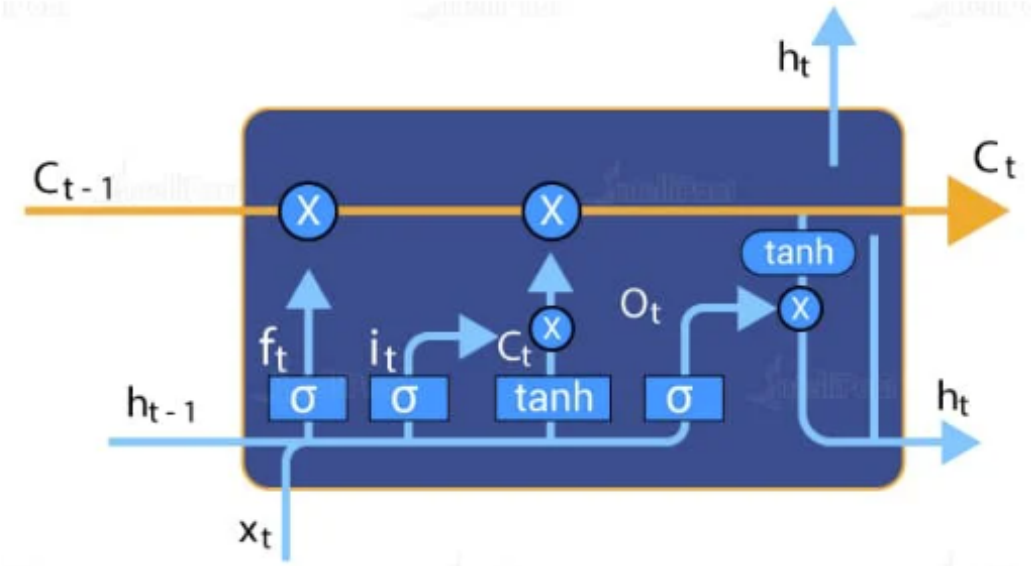

Figure 1. Internal structure of the LSTM.

LSTM has three gates to control the storage state, including the forget gate, the input gate and the output gate.

- **Input gate:** It determines which of the input values should be used to change the memory. The sigmoid function determines whether to allow 0 or 1 values through. And the tanh function assigns weight to the data provided, determining their importance on a scale of -1 to 1.

$$i_t = \sigma(W_i . [h_t - 1, x_t] + bi)$$
$$C_t = \tanh\,(W_C . [h_t - 1, x_t] + b_C)$$

- **Forget Gate:** It finds the details that should be removed from the block. It is decided by a sigmoid function. For each number in the cell state Ct-1, it looks at the preceding state (ht-1) and the content input (Xt) and produces a number between 0 (omit this) and 1 (keep this).

$$f_t = \sigma(W_f . [h_t - 1, x_t] + b_f)$$

- **Output Gate:** The block's input and memory are used to determine the output. The sigmoid function determines whether to allow 0 or 1 values through. And the tanh function determines which values are allowed to pass through 0, 1. And the tanh function assigns weight to the values provided, determining their relevance on a scale of -1 to 1 and multiplying it with the sigmoid output.

$$O_t = \sigma(W_o[h_t - 1, x_t] + b_o$$
$$h_t = o_t * \tanh(C_t)$$

3. Experiment:

3.1 Dataset

The data set adopted in this paper is the stock change information of Agilent stock (A ticker), including the stock trading information of the three years from January 1, 2011 to January 1, 2022. The data is divided into 80% training set, and 20% test set. The model is trained using the training set, and the hyperparameter is adjusted and the performance of the model is tested using the test set.

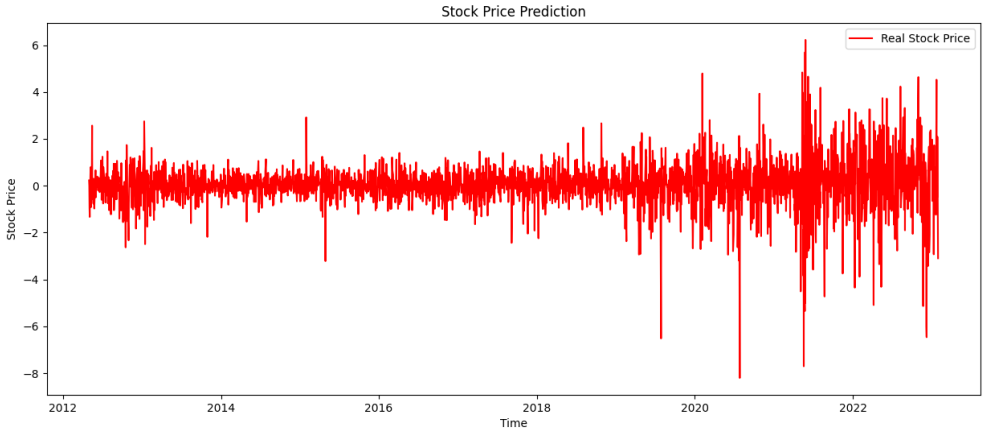

Figure 2. Agilent (A) stock price.

3.2 Evaluation Indicators of the Model

In this paper, Root Mean Square Error (RMSE) is selected to quantitatively evaluate the performance of the LSTM model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$$

yi refers to the forecast data of the stock price and y hat refers to the real data of the stock price, the smaller error between the predicted and real data of the stock price, and the better the performance of the model

3.3 Experimental Result and Analysis

The RMSE is used as model evaluation indicators.

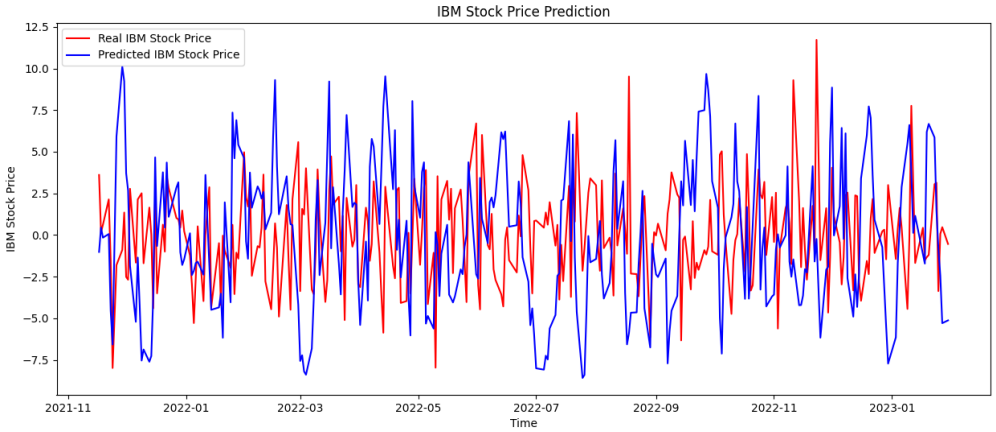

Figure 3. Agilent (A) real vs forecasting stock price

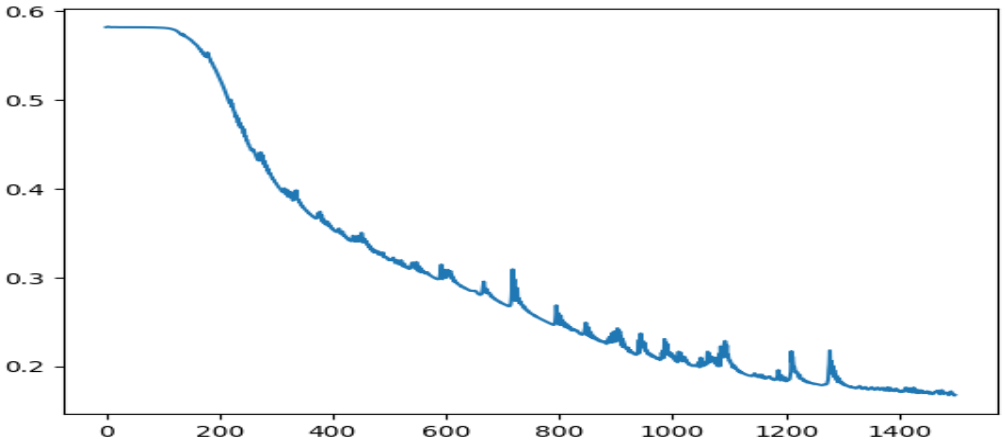

Figure 4. RMSE training loss graph

Table 1. Comparing the stock price prediction between train dataset and test dataset

	Training Set	Testing dataset
RMSE	0.54	5.09

4) Affected Stock Company analysis:

Stocks can be positively correlated when they move up or down in tandem. A correlation value of 1 means two stocks have a perfect positive correlation.If one stock moves up while the other goes down, they would have a perfect negative correlation, noted by a value of -1. If each stock seems to move completely independently of the other, they could be considered uncorrelated and have a value of 0.

4.1) Correlation method:

Correlation Coefficient is a statistical concept, which helps in establishing a relation between predicted and actual values obtained in a statistical experiment. The calculated value of the correlation coefficient explains the exactness between the predicted and actual values. Correlation Coefficient value always lies between -1 to +1. If correlation coefficient value is positive, then there is a similar and identical relation between the two variables. Else it indicates the dissimilarity between the two variables.

$$\rho(X,Y) = E\frac{(X-\mu_x)(Y-\mu_y)}{\sigma x.\sigma y}$$

4.2) Experimental result and Analysis

Find most affected 5 companies which related to Agilent (A) stock change using heat map and their correlation with each other. Five most affected companies are the top five highest correlations which respect Agilent (A) stock. This study proposes a Sort Algorithm which take maximum O(nlog(n)) and n less or equal than 500, n is the number of desired relating companies.

5 most affected companies which greatly affected by Agilent (A) stock change:

- Mettler-Toledo International Inc (MTD)
- Waters Corporation (WAT)
- Thermo Fisher Scientific Inc (TMO)
- SPDR S&P 500 ETF Trust (SPY)
- PerkinElmer, Inc (PKI)

5) Conclusion:

In this study, the research method of stock price time series prediction based on the LSTM model is proposed. Taking the stock price change of Agilent as an example, standardization is used to make the dataset more stable which can be affected by real factors to the stock price, and LSTM is used to predict stock price. The experimental data showed that the RMSE of LSTM is low which can prove that the LSTM architecture can learn the features efficiently. Furthermore, the study also finds out the relation of stock companies, and finds out most affected companies when stock prices of a company change. The following work will try to make larger data, further optimize the model and parameters, and improve the performance of the model.