Prediction of Facebook and GOOG Prices based on Linear Regression and LSTM Regression

Xiaonan Zuo*

School of Beijing University of Technology, Beijing, 100000, China

*Corresponding author email: xiaonan.zuo.@ucdconnect.ie

Abstract. Stock market analysis is a very difficult task, and stock markets are very complex and constantly changing environments. More and more stock investors are now becoming aware of the prominence of machine learning in the field of stocks and finance, and over the last decade or so machine learning has driven advances in the stock market, such as the ability to use different machine learning methods to predict stock movements in order to arrive at the best decisions and algorithmic trades. The problem that this project wants to investigate is the use of machine learning methods for stock prediction. Two stocks, Facebook and GOOG, were chosen as the datasets for the study. The datasets consisted of stock information from the last decade or so and two machine learning methods, namely long and short term memory and linear regression, were used to make predictions. The results obtained from these two models were analyzing and different results were obtained. The results present the conclusion that the linear regression model is more suitable than the LSTM model for predicting these two groups of stocks. Some error analysis was also carried out and some improvements were given for the two different models.

Keywords: Long-short term memory; linear regression; stock prediction; machine learning.

1. Introduction

In recent years social media has gradually taken its place in people's lives and more and more people are willing to share themselves on social media, which adds a lot of colour to our lives. Social media has given us the platform to share with others. Positive things can make people more passionate about life and convey positive energy. Facebook is an American social networking service website founded on February 4, 2004.

Facebook has not been available in China, so ordinary people are not familiar with it, but Facebook's global status is approximately equal to that of WeChat in China [1]. Tencent only occupies the social market in China and its Market capitalisation has reached 3 trillion RMB. Facebook's company (or meta as it is called) also has Instagram and WhatsApp under its umbrella, so imagine the feeling of being left out of the world if you don't use any social media all day, so whether active or passive, everyone's social network has been moved to platform on top of it, and its stock is even more relevant for research and mining potential [2, 3].

On the other hand, browsers are also inseparable from people's lives. Google search engine is currently the best and largest in the world. It has an advantage in terms of information and news, and Google has many useful tools [4, 5]. (Google has many useful tools (such as maps, translations, trend analysis, etc.).

Google is also a very representative one. According to foreign media reports, Google was once the most popular growth stock on the market [6]. Considering Google's strong growth prospects for the next few years, its share price is already quite cheap compared to its other large-cap, growth tech peers [7]. Therefore, it is also of great research interest to compare Google with Facebook to predict the direction of both stocks and to summarize the future of both stocks [8].

Stock forecasting is still in a hot phase in the last few decades, with stock markets in a volatile situation and with many external factors that can affect stocks and make them fluctuate significantly, making it difficult for shareholders to invest profitably in stocks [9,10]. Theoretically, we can use machine learning methods to forecast stocks, and also the volatility of stocks is influenced by many factors, but the impact of different factors on stocks is not clearly defined yet. Stock forecasting is usually non-linear, so the choice of model needs to be such that the model can handle non-linear

forecasting problems and, because of the time-series nature of stocks, recurrent neural networks are suitable, for stock forecasting.

2. Method

2.1 Long Short Term Memory (LSTM)

For the choice of model, both linear regression and LSTM methods were chosen for this project [11]. The data fed into the LSTM network must be in the format (number of samples, time step, number of features), so we need to process the data a little. One thing to note here is that it needs to be normalised, otherwise convergence is difficult.

The performance of traditional machine learning algorithms (linear or logistic regression, SMV, random forest, etc.) increases as the training dataset grows, but after a certain point, the performance of the algorithm stops increasing. After the dataset size exceeds this value, even if more data is provided to the model, the traditional model does not know how to handle the additional data and thus performance does not improve further. This is not the case with neural networks, where this never happens. The performance of a neural network always increases as the amount of data increases (of course: this data is good quality data, and as the size of the network increases, the speed of training increases. Therefore, if you want to get the best performance, you need to be somewhere on the right side of the x-axis (high data volume) on the green line (large neural network).

The principle of time prediction is to extrapolate the results of N+1 moments from the previously known results of N moments by processing the results analytically. This type of learning supervision problem is very typical in machine learning. For its variables, the output is taken as the unknown required prediction result and the input as the result that existed in the past. When using the LSTM model for data prediction, the data in the given datasets will be the first processed to transform the original data into a type of data that can be applied in machine learning, where one row can store one sample, i.e. The number of samples is the same as the number of rows, and the variable book is the same as the number of columns.

LSTM memory is a model that increases the memory of recurrent neural networks. Recurrent neural networks retain short-term memories because they allow information to be determined earlier in the current neural network. For immediate tasks, RNNs use earlier data, but we may not be utilising all the earlier information of the neuron. LSTMs are widely used in RNNs. One of the main problems with RNNs is the gradient disappearance problem, which arises from using the same parameters repeatedly in RNN blocks. We must try to overcome this problem by using different parameters at each time. Early on, time series data were treated as static, such as the opening and closing price of a story [12].

The LSTM is formed by a chain of repeating modules, each individual module is called a cell (Metacell). Each Metacell has a specific gate structure to selectively allow messages to pass through, and state updates for each Metacell are achieved by message passing through the LSTM gate structure.The structure of the Metacell is shown in Fig 1.

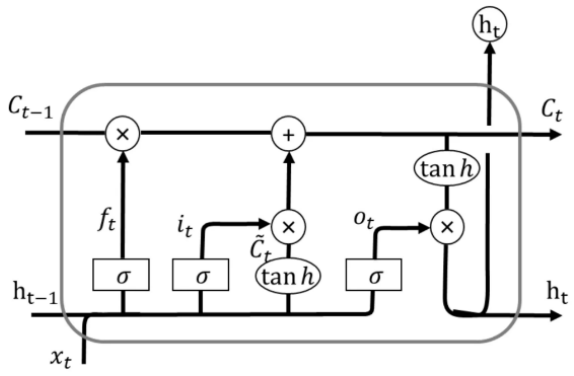

Figure 1. LSTM process image

Where $\sigma$is the Sigmoid activation function, denoting the basic product operation (Hadamard), and tanh denotes the tanh activation function.

The first step in the LSTM is to determine the update of the metacellular state. It processes (previous output) and (current input) through the Sigmoid function, outputting a number from 0 to 1. 1 means all retained, while 0 means all discarded. Where w_f is the weight matrix and b_f is the bias term.

$$f_t = \sigma(\, W_f \times [h_{t-1}, x_t] + b_f) \tag{1}$$

Input gates are used to input data, and the inputs determine the state of the message. The first step is the Sigmoid layer, which stores the weights of the updated input information. i.e. those messages need to be updated. Next candidate vectors are formed from this calculation by tanh construction $\tilde{C}_T$.

$$f_t = \sigma(\, W_f \times [h_{t-1}, x_t] + b_f) \tag{2}$$

$$\tilde{C}_T = tanh\ (W_c \times [h_{t-1}, x_t] + b_c) \tag{3}$$

Next the previous cell state C_{t-1} is updated and this matrix is added to the previous meta-cell state after the forgetting gate process to form C_t. Multiplying the previous state matrix by f_t represents the selective memory of the previous cell state. Afterwards we add the resulting value to $i_t * \tilde{C}_T$ to obtain the new cell state C_t.

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_T \tag{4}$$

The final output gate will be based on the current cell state, passing the current input data through the Sigmoid layer. The current cell state is then passed through the tanh layer, which normalises the matrix to between -1 and 1 and performs a basic product operation with the output of the Sigmoid layer, to this point, determining the data for the output section.

$$o_t = \sigma(\, W_0 \times [h_{t-1}, x_t] + b_0) \tag{5}$$

The modular chain network of LSTM cells, as described above, enables predictions to be made about the future through continuous learning of the input data.

$$h_t = o_t \times tanh\ (C_t) \tag{6}$$

For linear regression, a method of machine learning, the model is based on a pair of two variables, the independent variable (x) and the dependent variable (y). In the context of multiple linear regression, there may be many independent variables. A simple linear regression has only one independent variable x. In the conditions given by the current model, i.e., the data set, there is only one independent variable, date. The first date that rises to the length of the date vector is represented by the integer 1. The length of this vector must be an integer and the date variable will change according to the time and, at the same time, the price of the stock, which is the dependent variable, will also change.

$$y = bx + a \tag{7}$$

$$Y = predicted\ value\ or\ dependent\ variable \tag{8}$$

$$b = slope\ of\ the\ line \tag{9}$$

$$x = coefficient\ or\ dependent\ variable \tag{10}$$

$$a = y - intercept \tag{11}$$

In OLS, we plot a large number of linear data on the datasets so that we can find the best fit that minimums the sum of the squared errors (SSE) between the predicted values from linear regression and the true values in the original datasets. For each line we predict and plot, there must be some error between the simulated prediction and the true value in the datasets. The sum of squares is obtained by comparing the predicted and true values, taking the square of each difference and adding the sum of squares.

2.2 Linear Regression

Linear regression models are classical statistical models, which are applied in scenarios where a continuous numerical variable (dependent variable) is predicted from a known variable (independent variable).

When we initially build a model, if there are some uncorrelated predictors in the p predictor variables, then we should select a subset of the better predictors from the middle to refit the model so that the predictive power and explanatory power of our model can be improved. By retaining some of the predictor variables and discarding the remainder, the subset selection results in an explainable model, and the subset selection results in an explainable model where the more refined model will have a higher prediction error than that model. A linear regression model can be a discrete based process where the variables are given two choices, choose to keep them or choose to drop them. Because of its high variance, it is concluded that the prediction error of the model is usually higher. On the other hand, the effect of compression is always present and therefore the effect of high variability is less.

Multiple linear regression considers a two-variable model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{12}$$

Here Y is the dependent variable, $X1$ and $X2$ are the independent variables or predictors, ε is the error term, and $\beta 0$, $\beta 1$, and $\beta 2$ are the unknown coefficients. Multiple linear regression, typically, where the dependent variable of the model is Y, is a linear function of k independent variables, the k independent variables being: $X1$ to Xk,

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon \tag{13}$$

When measuring multivariate models, the original R_2 metric is no longer appropriate as the addition of additional variables increases R_2, hence the concept of adjusting R_2. i = 1 when there is an intercept term and vice versa equals 0. n denotes the number of predicted outcomes under the fitted model, p represents the number of parameters required to fit the model. Note that adjusting R_2 is only used for model selection and the original R_2 is used when the explanatory power of the model is to be stated.

Linearity, a linear relationship between the dependent and independent variables, means that the model parameters and the explanatory variables are linearly related to each other, and that the explanatory and explanatory variables can be arbitrarily related to each other and can be transformed by any function before regression. The orthogonality assumption is that the error term is not correlated with the independent variable and its expectation is 0.

$$E(ux) = 0 \tag{14}$$

i.e. the nuisance term and the explanatory variable cannot be linearly related. This assumption suggests that when building a model, any variable that is correlated with both x,y should be included in the model, otherwise the regression coefficients are biased. This assumption cannot be tested on the basis of the results after the regression because the least squares method is itself an orthogonal transformation and even if this assumption is not satisfied, any estimation method that produces residuals will be orthogonal to the explanatory variables.

When it comes to linear regression, the first thing we need to remember is the concept of correlation analysis, which means that before conducting a regression analysis, we generally conduct a correlation analysis to determine whether the dependent and independent variables are indeed related (this can be seen as a form of exploratory data analysis). General we can verify the correlation between the independent and dependent variables by means of a scatter plot, and if there is a clear linear relationship, the regression can be carried out with confidence. If there is a non-obvious linear relationship, the analysis can only continue with a certain amount of data variation, or perhaps no pattern can be seen between x and y, and then regression analysis cannot be performed at that point.

When it comes to correlation, in general Pearson's correlation coefficient can be used to indicate the magnitude of the previous correlation of the data, which is a value between -1 and 1. The closer it is to 0, the less correlated it is, and conversely the more correlated it is. We can also perform a test of correlation by representing the overall correlation coefficient by ρ, which is estimated from the sample statistic r. The null hypothesis H_0: $\rho = 0$, if the original hypothesis is rejected: ρ the correlation coefficient is not 0. The Pearson correlation coefficient is a parametric method used to determine the correlation between continuous variables. The latter two non-parametric methods can be used when the data do not satisfy normality, while the Spearman correlation coefficient, also known as the rank correlation coefficient, can be used to evaluate monotonic relationships before coefficients. Kendall correlation coefficient is a rank correlation coefficient, which is calculated for categorical variables. The objectives of linear regression are the following: to assess the significance of the variation and performance of the response variable in the analysis of the variables obtained from the prediction; and to predict the outcome of the corresponding vector given the outcome of the predictor variable. Consider a model with one variable:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{15}$$

Where Y is the dependent variable (in this case the mean of Y), X_1 is the independent or predictor variable, ε is the disturbance term and β_0, β_1 are unknown coefficients. The least squares method (OLS) is commonly used to estimate β_0, β_1, and OLS is the optimal linear unbiased estimate.

3. Results and Discussion

The general approach of these studies is to predict future returns from existing data using LSTM algorithms and linear regression, then use thresholds to determine upward or downward transitions and assign equal weights to assets with an upward predictive direction. The main objective of the experiment is to predict the closing price of a stock at a certain point in the next trading day in a fairly short period of time, and the selected characteristics are the opening price, closing price, minimum price, maximum price, trading volume and trading volume. Analysis of the prediction outputs showed that the machine learning model was effective in predicting the future closing price trend of the stock in the short term.

As can be seen from the evaluation metrics, there are no significant inaccuracies in the results obtained using these two methods, and we also visualise the output of the results. The predicted fit is basically the same as the actual true value of the trend, but it is still not identical and there is a certain amount of error, which needs to be improved. The direct output of the values gives a more intuitive comparison of the differences between the two models, with mean square error, root mean square error, mean absolute error and R-square being output. Overall, the LSTM model and the linear

regression model are effective in predicting stock prices and are useful as a guide for the securities investment market.

Facebook stock, as figure 2 shows:

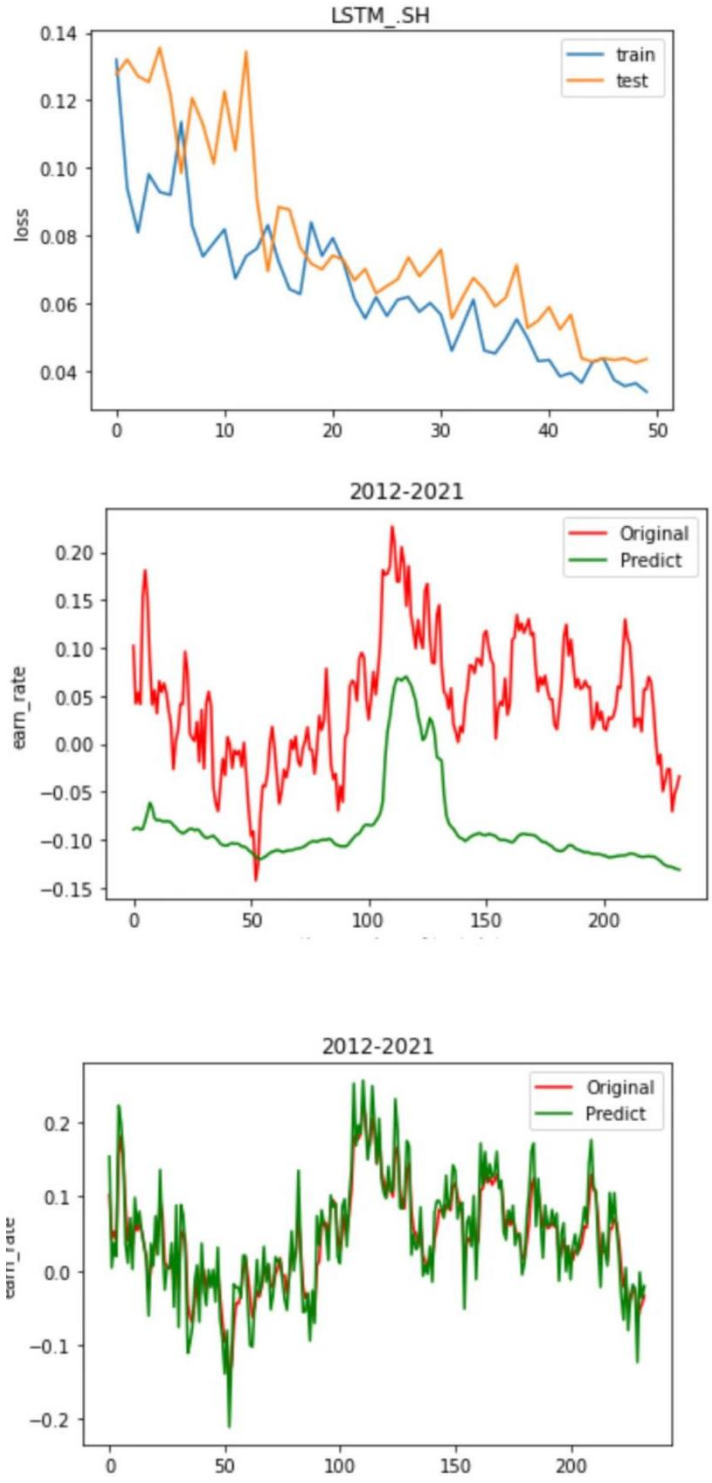

Figure 2. Facebook stock result

GOOG:

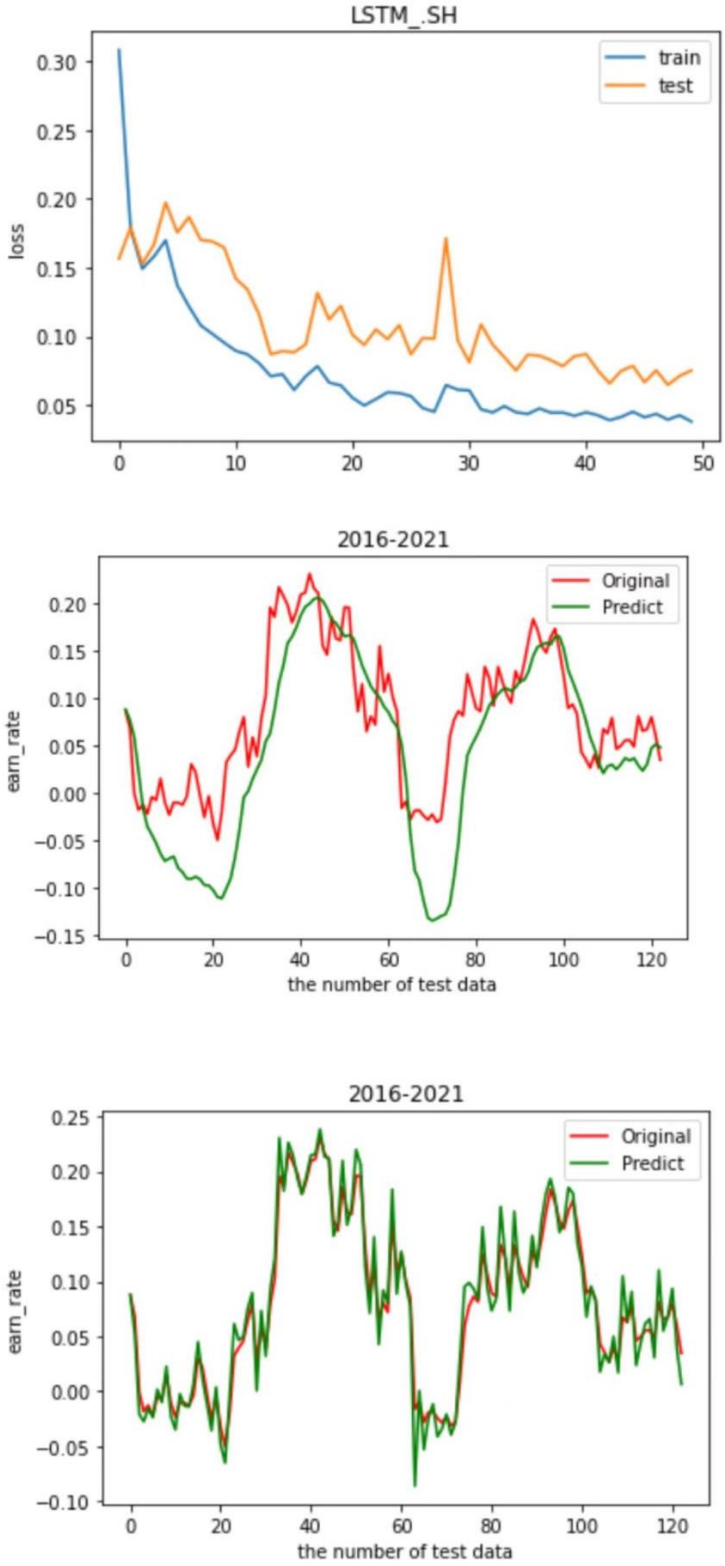

Figure 3. GOOG stock result

As figure 3 shows, the statistical parameter presented below is the average of the squared error between the predicted result and the corresponding point of the input value, calculated as SSE/n. SSE means sum of squared errors, like the difference between the true and predicted results is then squared, and MSE is defined as SSE/n. The smaller the SSE and MSE, the better the fit of the model and the more accurate the prediction. MSE for both data sets is smaller than the LSTM fit for the linear

regression fit. RMSE is a statistical parameter that is the square root of the MSE, again the smaller the value of RMSE, the better the model fits, and at this point the MSE of both linear regression fits is smaller than the MSE of the LSTM fit. It is a data criterion that characterises the degree of fit. The closer it is to 1, the more accurately the variables in the equation predict the y-value and the better the model fits that data set. The R-squares fitted by linear regression are all closer to 1 than those fitted by LSTM. For both sets of data, all three characteristic quantities indicate that the linear regression model is a better fit than the LSTM model.

4. Conclusion

The stock market is a multivariate non-linear dynamic system for which there are currently no better academic modelling methods. At the same time, the stock market has a degree of uncertainty and therefore the prediction results of those methods that attempt to build an exact model for stock price forecasting are bound to be unsatisfactory; the non-linear nature of the stock price itself and the variables that affect it require a strong ability to deal with non-linear.

The complexity of the correlations between various factors in the stock market, the variability of primary and secondary relationships and the difficulty of extracting quantitative relationships, make it very difficult to apply conventional forecasting methods to quantitative analysis of the stock market. There are many stochastic factors in the stock market, which have a significant impact on stock indices and prices, with dramatic price fluctuations and high dryness, showing a strong non-linearity and uncertainty. Many methods have been proposed for stock market forecasting at home and abroad. These methods have an important guiding role in practical work, but there are still certain unsatisfactory aspects, such as poor extrapolation of regression models, poor accuracy of the analogous coefficient method, and the large computational effort of neural networks and their tendency to produce over-fitting.

References

[1] Yoo S, Jeon S, Jeong S, et al. Prediction of the Change Points in Stock Markets Using DAE-LSTM. Sustainability, 2021.

[2] Borovkova S, Tsiamas I. An Ensemble of LSTM Neural Networks for High-Frequency Stock Market Classification. Social Science Electronic Publishing, 2020.

[3] Livieris I E, Pintelas E, PintelasP. A CNN–LSTM model for gold price time-series forecasting. Neural Computing and Applications, 2020, 1 - 10.

[4] Li Jiaying, et al. A variable weight combination model for prediction on landslide displacement using AR model, LSTM model, and SVM model: a case study of the Xinming landslide in China. Environmental Earth Sciences, 2021, 80 (10): 1 - 14.

[5] Xin Y, Xiao G S. Linear Regression Analysis: Theory and Computing. World Scientific, 2009.

[6] Laumann, Yanina. Analysisof beta coefficients in the brazilian stock market using fuzzy linear regression methodology. Fuzzy Economic Review, 2015, 20: 3 - 17.

[7] Alkhatib K, et al. Stock Price Prediction Using K-Nearest Neighbor Algorithm. Ijbhtnet Com, 2013.

[8] Maguluri L P, Ragupathy R. A Cluster based Non-Linear Regression Framework for Periodic Multi-Stock Trend Prediction on Real Time Stock Market Data." International Journal of Advanced Computer Science and Applications, 2020.

[9] Emioma C C, Edeki S O. Stock price prediction using machine learning on least-squares linear regression basis. Journal of Physics: Conference Series, 2021.

[10] Cosenza D N, et al. Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock. Forestry, 2020, 1 - 13.

[11] Lu W, et al. A CNN-LSTM-Based Model to Forecast Stock Prices. Complexity, 2020, 1 - 10.

[12] Chen K, et al. A LSTM-based method for stock returns prediction: A case study of China stock market. IEEE International Conference on Big Data IEEE, 2015, 2823 - 2824.