Financial Forecasting based on LSTM and Text Emotional Features

He Wang[1] ,Zhiqiang Guo[1]
1. School of Information Engineering, Wuhan University of Technology
Wuhan, Hubei Province, China
wh_2012@whut.edu.cn, guozhiqiang@whut.edu.cn

Lijun Chen[1]
1. School of Information Engineering, Wuhan University of Technology
Wuhan, Hubei Province, China
chenlj@whut.edu.cn

***Abstract*—In order to realize the prediction of financial time series, this paper first collects relevant text information on Sina Finance, and extracts and obtains the time series of financial sentiment index with three indexes through sentiment analysis of the text. At the same time, this paper collects the time series of nine financial indices formed by the Shanghai Stock Exchange Index (SSE), and combines the two as the research object. Considering the long-term delay of financial time series, this paper chooses long short term memory (LSTM) neural network to construct prediction model. After training and testing the model, the results show that after increasing the text sentiment index, The mean square error (MSE) of the prediction results of the LSTM prediction model decreased from 74.57 to 19.06, and the mean absolute error (MAE) decreased from 5.96 to 3.14, and with the cyclic neural network (RNN) and the particle swarm back propagation neural network (PSO-BP). The prediction results are more accurate than the prediction model. The error is small and the accuracy is improved.**

Keywords: LSTM, time series, forecast, text emotion

I. INTRODUCTION

In modern economic activities, the securities market occupies an important position, affecting all aspects of the national economy and the people's livelihood, and it can reflect and influence the operation of the real economy. The prediction of the trend of the stock market index can not only obtain economic benefits, but also serve as the basis for decision-making by relevant policy makers. Therefore, it is of great significance for the forecast of securities market trends.

At present, the method of using the neural network and other methods to predict the financial index has been widely concerned. In [1], the least squares support vector machine is used to predict the financial time series, and the experiment is carried out on the Shanghai and Shenzhen 300 full-yield index, which has achieved good results. Literature [2] uses BP neural network to predict financial fluctuations. In [3], for BP neural network, the parameters of neural network are optimized by LM algorithm, which improves the accuracy of prediction. In view of the long delay in the financial time series, the traditional neural network will produce the characteristics of gradient explosion. The literature [4-6] uses RNN, LSTM and other cyclic neural networks as the prediction model to solve the above problems, using domestic and international securities. The market index experimented and achieved good results.

On the other hand, it has been found that by collecting and analyzing the emotional state of the public and investors, it is possible to predict the trends of relevant financial indices. The literature [7] uses the social media twitter to reflect the changes in public sentiment to predict the closing price. The literature [8] uses the web search intensity represented by Google Trends to predict the investor's reaction to the market, and the literature [9] passes the text. The analysis builds the manager's sentiment index, which is used to predict the overall return of the future securities market.

Based on the above background, this paper proposes a financial forecasting model based on LSTM neural network. Based on the existing financial index time series forecasting method, this model forms financial text by acquiring financial related texts and processing into emotional index data. The emotional index time series, which is combined with the financial index sequence to form a time series, thereby predicting the financial index.

II. LSTM NEURAL NETWORK MODEL

The original LSTM (long short term memory) long- and short-term memory neural network was proposed by Hochreiter et al in 1997, and then improved by Gers in 2000 to form the current LSTM neural network model. It is a structurally special RNN (Recurrent Neural Network) cyclic neural network. By using memory blocks instead of hidden layers, input gates, output gates, and forgetting gates are used to improve the network structure, so that LSTM can not only be like RNN. The same use of internal memory to solve problems, and can forget the useless information in the process of communication, thus solving the gradient disappearance of RNN in the face of long-term dependence problems. LSTM has a better effect than other neural network models in the processing sequence affecting longer delays. It is widely used in various fields such as text sentiment analysis, financial time series prediction, and text translation. Its network topology is shown in Fig.1.

This research was supported by the National Natural Science Foundation of China (No. 51879211).

978-1-5386-8178-7/19/$31.00 ©2019 IEEE

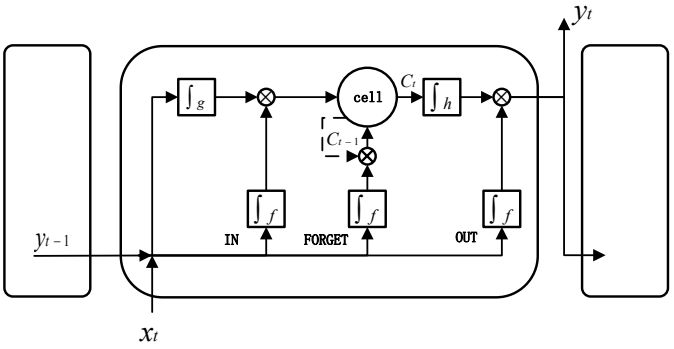

Fig. 1. LSTM neural network structure topology.

As shown in the figure, the LSTM neural network consists of a number of repeated structures, in each structure, consisting of an input gate, an output gate, a forgetting gate, and a memory cell. The input at time t is x_t, the output is y_t, the state of the memory unit is C_t, and the outputs of the input gate, the forgetting gate, and the output gate are respectively i_t, f_t and o_t.

At the input layer, both x_t and y_{t-1} are received, and the activation function (usually taking the tanh activation function) is input as time t, ie:

$$g_t = tanh(\mathrm{x}_t, y_{t-1}) \tag{1}$$

At the input gate, x_t and y_{t-1} are also received, and a vector of 0-1 is formed by the activation function (usually taking the simgoid activation function), namely:

$$i_t = \sigma(\mathrm{x}_t, y_{t-1}) \tag{2}$$

The vector is supplied to the memory unit after being multiplied by g_t. This embodies the filtering and control of the input of the input gate to the time t.

The forgotten gate also receives x_t and y_{t-1}, a 0-1 vector formed by the activation function:

$$f_t = \sigma(\mathrm{x}_t, y_{t-1}) \tag{3}$$

Multiply the memory cell state C_{t-1} at time t-1 and then return to the memory cell. In this process, the forgetting gate realizes the memory and forgetting control function of the memory unit for the t-1 time state.

The state of the memory unit at time t consists of the above two parts:

$$C_t = g_t \cdot i_t + f_t \cdot C_{t-1} \tag{4}$$

It contains both the state of the input value at time t and the state of the memory cell at time t-1.

Finally, at the output gate C_t passes the output value of the activation function:

$$\mathrm{h}_t = tanh(C_t) \tag{5}$$

The 0-1 vector obtained with the output gate:

$$o_t = \sigma(x_t, y_{t-1}) \tag{6}$$

Click multiply to get the output value at time t:

$$y_t = \mathrm{h}_t \cdot o_t \tag{7}$$

Based on the above formula, the output y_t at time t can be obtained from the output y_{t-1} at time t-1 and the input x_t at time t.

III. PREDICTION MODEL BASED ON LSTM AND TEXT SENTIMENT FEATURES

A. Time series construction

In order to implement the prediction model, it is first necessary to construct a corresponding time series data set to form a training set and a test set. For the purposes of this paper, the construction of this time series relies on financial time series based on financial indicators and emotional index time series based on sentiment analysis results of text data.

1) Text data acquisition

In order to ensure the correlation between the text data used for forecasting and the financial index, this article selects the Sina Finance (https://finance.sina.com.cn/) website as the source of text data. The related articles of the website are related to the financial data. Strong correlation and can reflect the investor's emotional tendency. According to the literature [10], the title can reflect the emotional tendency more strongly than the body content, so the title of the article is stored separately. Therefore, this paper designs and implements a web crawler, crawls each article page, and obtains three data: title, content text, and publication date. The body content is obtained by filtering the content text. Finally, the two items of the title and the body are used as the original text data, and are divided according to the date of publication.

2) Emotional index acquisition

In order to transform the text into the corresponding emotional index, this paper selects the open source Chinese text analysis system NLPIR to analyze and process the collected article title and body text, and obtain the corresponding emotional index representing the positive and negative proportion of the text content. Between 0 and 0, 0 represents complete negative emotions and 1 represents completely positive emotions.

After all the texts are divided according to the date, the sentiment index corresponding to the multiple articles under each date is calculated, and the mean value of the emotional index of the date is obtained, and the average of the two items of the article title emotion index and the body emotion index forms a time series. Text sentiment index time series.

3) Financial data acquisition

Considering that the text data comes from the Chinese portal website, the Shanghai Stock Exchange's Shanghai Stock Exchange Index (SSE) is selected as the source of the financial index data. The Shanghai Composite Index contains nine financial indicators that reflect the market's changes, such as closing price, highest price, lowest price, opening price, previous closing, ups and downs, ups and downs, volume, and transaction amount. According to the chronological order of the index, a time series containing nine financial indices is formed.

B. LSTM predictive model construction

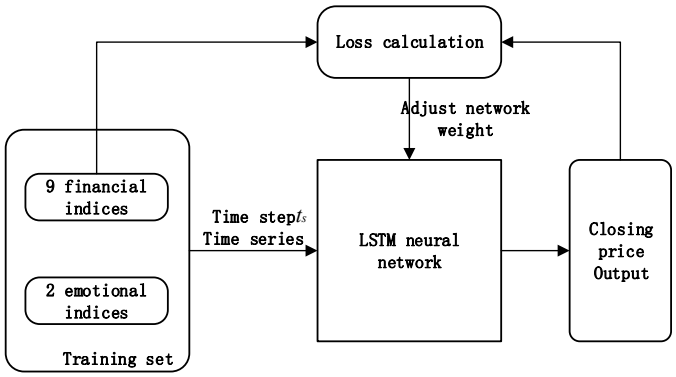

Fig. 2. LSTM prediction model.

As shown in Fig.2. the LSTM prediction model is constructed for this paper. In this model, nine financial indexes and two emotional indices are used as training sets for the model. When the model is training, a total of 11 parameters are used as input data. Considering that the key to the prediction of financial indicators is to predict the future closing price, the closing price is selected as the output of the forecasting model.

The steps to train the model using the training set are as follows:

1. Set the number of time steps t_s, learning rate l_r and other parameters,

2. Enter 11 indexes in the training set from 1 to $t_s$days.

3. Calculate the closing price of the corresponding result by predicting the model.

4. Calculate the loss based on the actual closing price of the t_s+1 day and the predicted closing price.

5. Using the back propagation algorithm to adjust the network weight,

6. Repeat 2-5 items. When the nth training is performed, input n to n+t_s days of data until all training sets are completed.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The Shanghai Stock Exchange's Shanghai Stock Exchange Index (SSE) was collected from January 2, 2014, February 1, 2018, with a total of 1,000 valid financial days, as the financial index data set. The page data on the Sina Finance website was collected, and through the text sentiment analysis, a text emotional index data set containing two indexes was obtained. Since the stock market will have intermittent rest days, the text sentiment index data set is processed according to the date label of the financial index data set, and part of the data is eliminated, so that the date labels of the two data sets are consistent, so that the two data sets are combined. The merger was performed to form a time series containing 11 indices.

The experimental environment of this paper is: Intel® Xeon E5-1620 CPU, clocked at 3.50GHz, 8GB memory, LSTM model is implemented in Python language version 3.6.5, using the relevant components of tensorflow 1.10.0 package, in RNN related model Based on the package, the prediction model of this paper is constructed. In the experiment, the number of time steps t_s was determined by repeated experiments, and the learning rate l_r was selected to be 0.0005. Therefore, the data set is divided by t_s=20, and the output of each group is the closing price of the 21st day, thereby constructing experimental data. 980 sets of experimental data were divided, 780 sets were used as training data sets, and 200 sets were used as test data sets.

In the experiment process, the model is trained first, and the network weight is adjusted by calculating the loss each time. After many experiments, the network weight with the smallest loss value is retained as the final prediction model.

As shown in Fig.3, for the loss of 1000 experiments, it can be seen from the figure that after 250 experiments, the loss remains basically unchanged, and the minimum loss value of 0.0013 appears in the 1802th experiment.

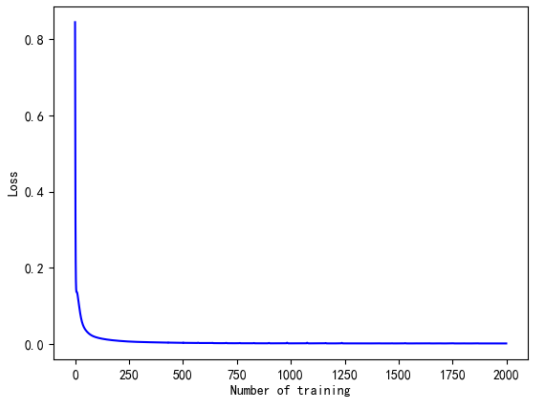

Fig. 3. Change in loss.

At the same time, this paper also constructs an LSTM prediction model that uses only nine financial index time series training. By comparing with the prediction results of the model, it can be used to test the influence of the text sentiment index on the prediction results.

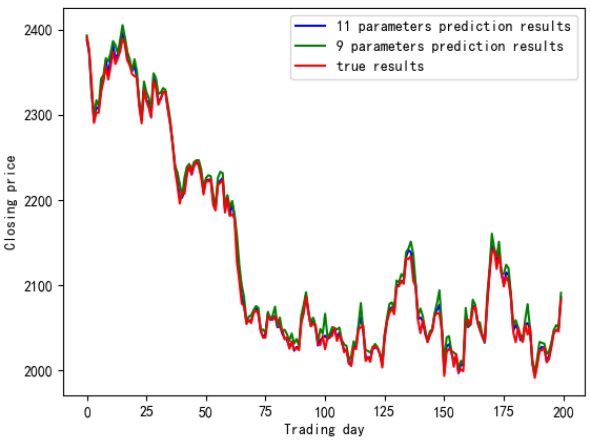

Fig. 4. Prediction results and true values of two models.

Fig.4 shows the comparison between the results of the 11-item index prediction model including the text sentiment index and the prediction model using the nine financial index prediction models. It can be seen from the figure that the results obtained by using only nine financial indices are relatively large, and the prediction model constructed in this paper has a relatively small fluctuation range and a more stable prediction result.

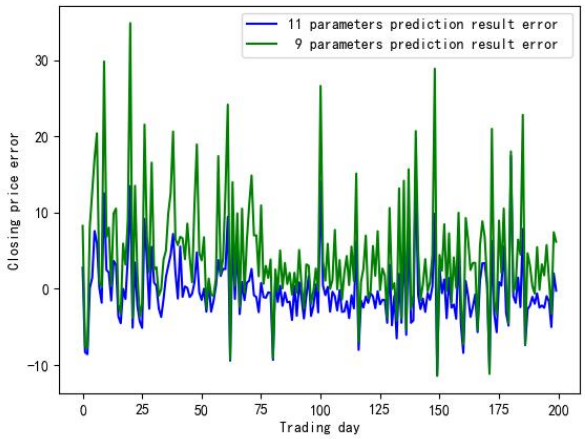

Fig. 5. Error value of the prediction result.

At the same time, according to the data in Fig.4, Fig.5 is calculated, which is the error value between the two prediction models and the true value. It can be seen from the figure that the results obtained by using only nine financial index predictions have larger wave errors, and the prediction models constructed in this paper have smaller errors and more accurate prediction results.

At the same time, based on the literature [11], a PSO-BP prediction model based on nine financial indexes is constructed, and a RNN prediction model is constructed. Based on the prediction data and real values of each model, the mean square error (MSE) and root mean square are calculated. The five index data of error (RMSE), mean absolute error (MAE), F value check, and P value check form Table 1.

TABLE I. TWO MODEL PREDICTION RESULTS INDICATORS

	MSE	RMSE	MAE	F	P
LSTM-11	19.06	4.37	3.14	0.02	0.97
LSTM-9	74.57	8.64	5.96	0.02	0.96
PSO-BP	159.32	12.62	7.48	0.04	0.79
RNN	93.17	9.65	6.21	0.03	0.85

It can be seen from the table that the prediction results of the sentiment index with the addition of two texts are smaller than the results of the prediction based on only nine financial indices, and the mean square error, root mean square error, and average absolute error are small. Explain that the text sentiment index can be used to predict the closing price, which can effectively reduce the error of the prediction result and improve the prediction accuracy. ,

In summary, after experimental comparison, the LSTM prediction model based on financial index and text sentiment index proposed in this paper is more accurate and stable than the prediction model based on financial index alone.

V. CONCLUSION

This paper proposes a securities market forecasting model based on financial index time series and text sentiment index time series, which uses LSTM neural network for prediction. The experimental results based on the Shanghai Stock Exchange's Shanghai Stock Exchange Index (SSE) show that compared with the traditional prediction method using only financial index time series, the model effectively combines the time series of text sentiment index, making the prediction result more accurate and error reduction. small. This paper verifies that the text sentiment index time series has certain significance for the securities market forecasting research, and illustrates the feasibility of studying the investor's emotional state and then predicting the market.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (No. 51879211).

REFERENCES

[1] Z. Y. Xin, G. Ming, “Comlicated financial data time serise forecasting analysis based on least square support vector machine,” JTsinghua Univ(Sci & Tech), vol. 48,issue 7, pp. 1147-1149, Jul. 2008.

[2] M. Xu, F. Wang, “Forecasting financial volatility based on BP neural network and symbolic time series,” Journal of WUT(Information & Management Engineering), vol. 37,issue 4, pp. 456-460, Aug. 2015.

[3] J. Xiao, Z. L. Pan, “Stock price short-time prediction based on GA-LM-BP neural network,” Journal of Computer Applications, vol. 32,issue 1, pp. 144-146, Jul. 2012.

[4] Chen K, Zhou Y, Dai F. A LSTM-based method for stock returns prediction: A case study of China stock market[C]// IEEE International Conference on Big Data. IEEE, 2015:2823-2824.

[5] Zhao Z, Rao R, Tu S, et al. Time-Weighted LSTM Model with Redefined Labeling for Stock Trend Prediction[C]// IEEE, International Conference on TOOLS with Artificial Intelligence. IEEE Computer Society, 2017:1210-1217.

[6] Selvin S, Vinayakumar R, Gopalakrishnan E A, et al. Stock price prediction using LSTM, RNN and CNN-sliding window model[C]// International Conference on Advances in Computing, Communications and Informatics. IEEE, 2017:1643-1647.

[7] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1):1-8.

[8] Dzielinski M. Measuring economic uncertainty and its impact on the stock market[J]. Finance Research Letters, 2012, 9(3):167-175.

[9] Jiang F, Lee J A, Martin X, et al. Manager Sentiment and Stock Returns[J]. Social Science Electronic Publishing, 2015.

[10] Mellouli S, Bouslama F, Akande A. An ontology for representing financial headline news[J]. Web Semantics, 2010, 8(2): 203-208.

[11] X. D. Miao, L. X. Wei, “Financial time series prediction based on wavelet analysis and PSO-BP neural network,” Information Techology, vol. 5, pp. 26-29, May. 2018.