2022 International Conference on Machine Learning and Cybernetics (ICMLC) | 978-1-6654-8832-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICMLC56445.2022.9941293

STOCK PRICE PREDICTION BASED ON LSTM AND BERT

XIAOJIAN WENG[1], XUDONG LIN[1], SHUAIBIN ZHAO[1]

[1]College of Mathematics and Information, South China Agricultural University, Guangzhou 510642, China
E-MAIL: weng450@163.com, hunanlxd@163.com, zhaoshuaibin1998@163.com

Abstract:

Price movements in the stock market affect all aspects of the social economy, and forecasting stock prices is of great importance. Traditional stock forecasting models are based on statistical regression models, which are difficult to characterize the influential relationships between multiple variables and predict stock price trends with large errors. In recent years, with the development of neural networks, neural networks have become a common method for stock forecasting, which include Back Propagation (BP) neural network, Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) neural network. However, most of the previous stock price prediction models only use the basic stock market data, ignoring the influence of stock market investor sentiment on stock prices. A new stock price prediction model is proposed to address the above problems. First, the investor sentiment before the stock opening is calculated by fine-tuning the BERT model, then the calculated investor sentiment and the basic stock quotation data are aggregated, and finally the LSTM model is used to predict the closing price of the next stock trading day. We validate the effectiveness of the model on a real dataset of three Chinese listed companies.

Keywords:

Long Short-Term Memory (LSTM); Stock prediction model; Investor sentiment; Machine learning

1. Introduction

The Stock market is the place where stock companies and investors trade shares. The stock market has existed for more than four hundred years and has become an indispensable part of the country's economy. On the one hand, the stock market is an effective means for joint-stock companies to gather capital from stock investors by issuing shares. By issuing shares, a large amount of capital flows into the stock market, promoting capital concentration, enhancing the organic composition of corporate capital, and greatly contributing to the development of the company's economy. On the other hand, stock investors, through their own judgment, choose companies that they believe have potential for development and invest their capital in them, helping the company to develop while also gaining wealth for themselves. Stock investors and joint stock companies complement each other, making the stock market a core driver of economic development.

However, the formation of stock prices is quite complex, including economic, political, market factors and investor behavior that can lead to changes in stock prices. The constant changes in stock prices provide room for speculative activities to survive and increase the risk of the stock market. This risk will not only bring economic losses to investors and stock companies but also may bring certain side effects to the country and society. Therefore, to make effective decisions on stock management and stock investment activities, investors need to analyze and grasp the potential development of the stock market and use it to forecast stock prices in order to avoid the risks associated with the stock market. Thus, stock price forecasting research can not only guide investors to make beneficial investments and thus earn personal income but also contribute to the development of the national economy.

With the advancement of mathematical theories and algorithms, a number of stock price forecasting models have emerged, which can be broadly classified into two categories based on different underlying theories: traditional time series forecasting models based on statistics and new time series forecasting models based on machine learning [1]. Autoregressive Integrated Moving Average (ARIMA) and Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) are two typical statistics-based time series models, and these types of forecasting models have played a role in stock forecasting [2-3], but the statistics-based time series models all use linear models to fit nonlinear nonstationary stock prices, which have many drawbacks, such as the difficulty of characterizing the relationship between multiple variables. Moreover, these methods often require some assumptions and pre-knowledge, such as the underlying data distribution and valid ranges for various parameters and their connections. However, as a complex system with many influential factors and uncertainties, the stock market tends to exhibit strong nonlinear characteristics, which makes the conventional analytical methods ineffective. In addition, the amount of information processed by the modeling and

978-1-6654-8832-7/22/$31.00 ©2022 IEEE

forecasting of the stock market is often very large, raising great challenges for the algorithm design. These characteristics make the prediction of the stock market based on conventional methods inadequate[4].

In recent years, new time series forecasting models based on machine learning and deep learning have started to emerge, and these new time series forecasting models, which can learn the relationship between multiple variables compared to traditional time series forecasting models, are more suitable for the stock market that changes from time to time. Xiao et al. [5] used support vector machines to forecast the China Shanghai Stock Exchange Composite Index, and the forecasting The results were better than the adaptive neuro-fuzzy system. Zhang et al. [6] used BP neural network to predict the closing prices of Gree Electric (000651) and BYD (002594) of Chinese A-shares, and the error between the predicted closing price and the real closing price was less than five percent in 17 out of 30 days less than five percent. The general neural network has the disadvantages of easily falling into local minima and only processing a single input, not extracting the relationship between the previous input and the next input, which is not conducive to the model to extract the stock market information closely related to time. LSTM is an improved RNN that solves the gradient disappearance and gradient explosion problems of general RNN to some extent. Chung et al. [7] used a hybrid model of LSTM and genetic algorithm to predict the Korean stock price index, and its prediction results outperformed the benchmark model. Compared with the conventional model, the LSTM-based stock price prediction model extracts the stock data features more adequately. However, these are not sufficient for stock price prediction.

In addition, economic, social and irrational human behavior and other factors can cause stock prices to move in ways that are not always in line with expectations. Investors' emotions, psychological states and behaviors are very important in the stock market. For example, Antweiler and Frank [8] found through experiments on email content and the Dow Jones index that email content helped predict stock market changes and confirmed the correlation between online comments and stock trading volume. In addition, reference [9] highlights the important role of emotions in investors' decisions. Although there are many studies showing a strong correlation between sentiment and stock prices, few studies have considered the use of sentiment analysis for stock forecasting, and this paper aims to contribute to this aspect.

In this paper, we derive the daily investor sentiment of the stock by collecting the daily opinions posted by investors about the stock on the Internet and analyzing their sentiment using the BERT model. The investor sentiment and the stock's opening price, closing price, lowest price, highest price, volume and amount traded are input as features into the LSTM, which ultimately predicts the closing price of the stock on the next trading day. The rest of this paper is organized as follows. In Section 2, we describe the structure of the long short-term memory network LSTM and the characteristics of BERT. In Section 3, we describe the collection of experimental data, the preprocessing of experimental data and the modeling process. Section 4 discusses the experiments and results, followed by conclusions in Section 5.

2. Related work

In this section, we will briefly introduce the LSTM neural network and BERT, which are used in the proposed method.

2.1. LSTM neural network

LSTM neural network was proposed by Hochreiter et al. [10] in 1997 and improved and extended by Grave [11]. LSTM is the improved RNN. By adding forget gate, input gate and output gate on the basis of RNN, the problem of RNN gradient disappearance and gradient explosion can be solved to some extent. The LSTM structure is shown in Figure 1.

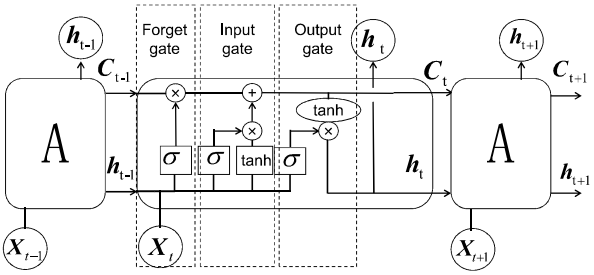


FIGURE 1. LSTM structure

The function of the forget gate is to decide which information from the previous moment is retained and which information is to be discarded. The process of the forget gate is:

$$\boldsymbol{g}_f = \sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{X}_t] + \boldsymbol{b}_f) \tag{1}$$

In formula (1), $\boldsymbol{W}_f$ and $\boldsymbol{b}_f$ are the weight and bias of the forget gate, respectively. The forget gate is used to control the memory state vector $\boldsymbol{C}_{t-1}$ at the moment of $t-1$ into the vector $\boldsymbol{g}_f * \boldsymbol{C}_{t-1}$ of the LSTM unit at the moment of t. The activation function σ is generally chosen as the Sigmoid function so that the value of $\boldsymbol{g}_f$ is between 0 and 1.

When g_f is closer to 1, the LSTM unit at the moment of t receives more information about C_{t-1}, and when g_f is closer to 0, the LSTM unit at the moment of t the less C_{t-1} information will be accepted.

The function of the input gate is that the information is selected and input to the input gate, and the input gate decides which information needs to be updated and how much. The process of the input gate is:

$$\tilde{C}_t = tanh(W_c[h_{(t-1)}, X_t] + b_c) \quad (2)$$

$$g_i = \sigma(W_i[h_{t-1}, X_t] + b_i) \quad (3)$$

Formula (2) is used to control the state h_{t-1} at time $t-1$ and the input vector X_t at time $t-1$ to obtain a new input vector $\tilde{C}_t$ through nonlinear transformation. Formula (3) is used to control the new input vector $\tilde{C}_t$ to enter the vector $g_i * \tilde{C}_t$ of the LSTM unit at time t. g_i also generally uses Sigmoid as the activation function. When the g_i is closer to 1, the LSTM unit at time t accepts more $\tilde{C}_t$ information. When the g_i is closer to 0, the LSTM unit at time t accepts less $\tilde{C}_t$ information.

The function of the output gate is: after the information is filtered by the forget gate and the input gate, it finally passes through the output gate to decide which information needs to be output. The process of the output gate is:

$$C_t = g_i * \tilde{C}_t + g_f * C_{t-1} \quad (4)$$

$$g_0 = \sigma(W_0[h_{t-1}, X_t] + b_0) \quad (5)$$

$$h_t = g_0 * tanh(C_t) \quad (6)$$

Formula (4) represents the update method of the memory state vector C_t of the LSTM unit at time t. Formula (5) and formula (6) represent the process of output gate, where C_t controls g_0 to generate output vector h_t through activation function $tanh$.

2.2. BERT

BERT (Bidirectional Encoder Representations from Transformers) [12] is a pre-trained language model based on Self-Attention, which completely dispenses with complexity and repetition, and its structure is shown in Figure 2. It is the Encoder part of the Transformer model [13] and consists of two subnets, namely the Self-Attention layer and the Feed Forward layer, each followed by a residual connection and a normalization layer. Due to the self-attention mechanism inside BERT, it is possible to train the bidirectional depth representation of the input text in parallel; moreover, the pre-trained BERT model can be followed by an output layer, which can be fine-tuned to fit the downstream task without a large amount of additional task-specific training, which can greatly reduce the training parameters and training time.

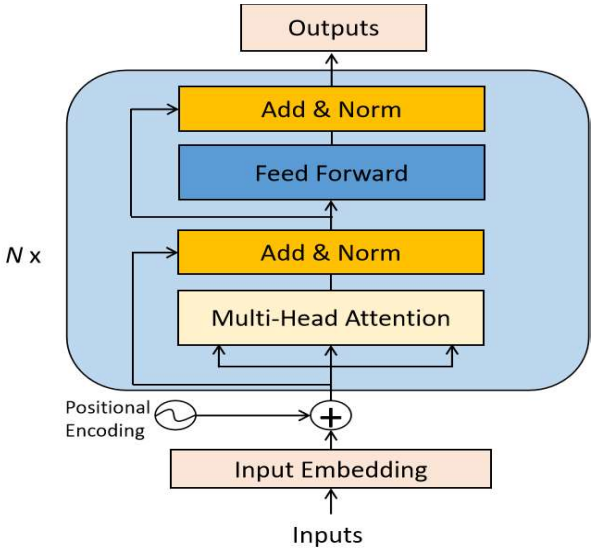


FIGURE 2. BERT structure

After BERT was proposed, it achieved extraordinary results on various tasks of NLP. There are three main reasons why the approach of fine-tuning the pre-trained BERT model can generalize well under different tasks as follows [14]:

1. Pre-training on the huge text corpus can learn universal language representations and help with the downstream tasks.
2. Pre-training provides a better model initialization, which usually leads to a better generalization performance and speeds up convergence on the target task.
3. Pre-training can be regarded as a kind of regularization to avoid overfitting on small data.

3. Methodology

3.1. Daily investor sentiment

From China's A-share PingAn Bank (000001), ZTE (000063), and MuYuan (002714), 9,000 post titles were randomly selected from each of the EastMoney Post Bars,

and investor sentiment was divided into three categories, namely positive, neutral and negative, with labels 1, 0 and -1 respectively. Download the Chinese BERT pre-training model from GitHub (https://github.com/google-research/bert), and use 7200 post titles as the training set, 900 post titles as the validation set, and 900 post titles as the test set for BERT fine-tuning. The learning rate is set to 2e-5, and the number of epochs is 3. The final classification accuracy on the test set is 66.56%.

Use the fine-tuned BERT model to classify all post titles, and then summarize the post titles according to the end of the previous stock trading day (3:00 PM Beijing Time) to the beginning of the next stock trading day (9:30 AM Beijing Time). Finally, stock investor sentiment for a given day is obtained according to formula (7).

$$sen_t = \sum_{h=1}^{l} neutral + \sum_{i=1}^{n} positive - \sum_{j=1}^{m} negative \quad (7)$$

In formula (7), sen_t is the investor sentiment on the t stock trading day, and $\sum_{h=1}^{l} neutral$, $\sum_{j}^{n} positive$, and $\sum_{j=1}^{m} negative$ are the total number of neutral sentiment post titles, the total number of positive sentiment post titles, and the total number of negative sentiment post titles after the close of day $t-1$ and before the open of day t, respectively.

3.2. LSTM implementation and configuration

The experiments in this paper are based on python 3.8, and the LSTM neural network is implemented based on the PyTorch deep learning framework. The model uses two LSTM layers, the number of neurons is set to 128, the optimizer is Adam, the learning rate is set to 0.001, the loss function is MSE, and the Timestep is set to 10. To prevent model overfitting and improve model performance, a dropout layer is added after each LSTM layer, and the dropout ratio is set to 0.2. 80% of the training samples are used for training and 20% for validating the model. The maximum training period of the model is set to 200, and the early stop method is added, so that the model will stop training when the metrics of the validation set do not improve in three consecutive training times.

3.3. Overall frame design

In summary, the stock price prediction model is divided into two parts, as shown in Figure 3. Sentiment analysis of daily investor sentiment is performed by fine-tuned BERT, and finally, investor sentiment and basic stock market data are added to the LSTM neural network to predict stock prices.

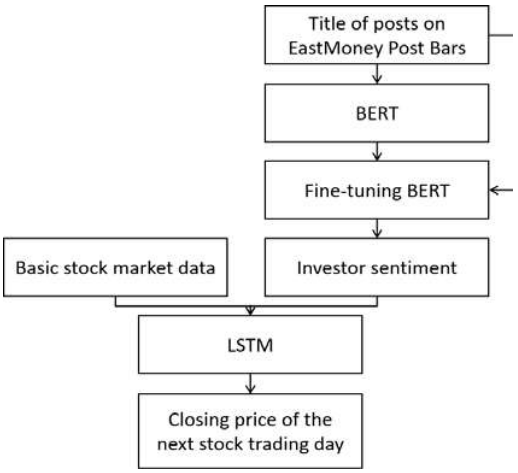


FIGURE 3. Model Framework

4. Experiment results and discussion

4.1. Data description

The experimental data used are PingAn Bank (000001), ZTE (000063) and MuYuan (002714) of China A-shares, from January 2, 2019 to September 24, 2021, for a total of 665 stock trading days, and the opening price, closing price, highest price, lowest price, trading volume and trading amount of each stock trading day are selected as the basic features. 85% of the data is used as the training set for training the model, and the remaining 15% is used as the test set. The text data for analyzing investors' sentiment were obtained from the post titles of the EastMoney (https://guba.eastmoney.com/), including 67,981 titles for PingAn Bank, 398,198 titles for ZTE and 109,956 titles for MuYuan.

4.2. Performance evaluation metric

To evaluate the validity of the model proposed in this paper, we use the following metrics:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - y_i\right| \quad (8)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{f_i - y_i}{y_i}\right| \quad (9)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2 \quad (10)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2} \quad (11)$$

$$Accuracy = \frac{\sum_{i=2}^{n} \mathrm{A}(\mathrm{A}(f_i - y_{i-1} > 0) = \mathrm{A}(y_i - y_{i-1} > 0))}{n-1} \quad (12)$$

Where f is the model prediction, and y is the true closing price of the stock. In formula (12), A is the indicator function; when the expression within the function is true, it is 1, and the expression within the function is false, it is 0.

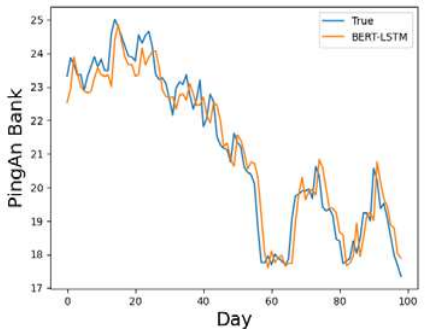
4.3. Experimental results and performance comparison

In this section, we refer to the proposed model as BERT-LSTM, and to verify the effectiveness of the model, we choose LSTM, Informer [15] with BERT-LSTM, using formulas (8), (9), (10), (11), (12) five evaluation metrics to compare their performance in the test set. The results are shown in Table 1.

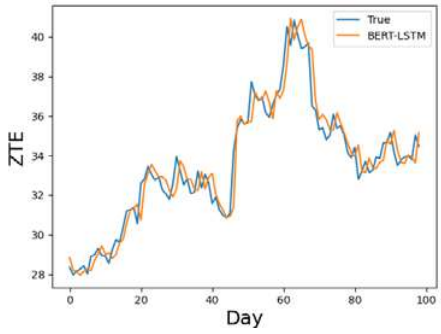
TABLE 1. Comparison of results of different models

Method		PingAn Bank	ZTE	MuYuan
LSTM	MAE	0.51	0.63	1.51
	MAPE(%)	2.42	1.87	2.83
	MSE	0.40	0.72	3.98
	RMSE	0.63	0.84	1.99
	Accuracy(%)	45.92	50.00	51.02
Informer	MAE	1.42	0.67	1.78
	MAPE(%)	6.36	2.02	3.25
	MSE	3.07	0.75	5.66
	RMSE	1.75	0.87	2.38
	Accuracy(%)	63.16	45.92	56.84
BERT-LSTM	MAE	0.46	0.59	1.33
	MAPE(%)	2.24	1.73	2.49
	MSE	0.36	0.67	3.29
	RMSE	0.60	0.82	1.81
	Accuracy(%)	46.94	61.22	57.14

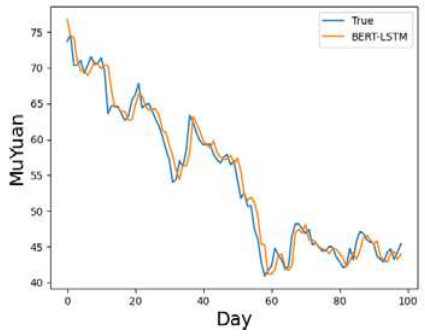
Carefully observed in Table 1, we found that in the BERT-LSTM, every evaluation indicator of each stock outperformed the other models except for PingAn Bank, whose Accuracy evaluation indicator was worse than Informer. Figure 4 shows the comparison between the closing price predicted by BERT-LSTM and the true closing price of the stock.



(A)PingAn Bank



(B)ZTE



(C) MuYuan

FIGURE 4. The three companies' closing price prediction results

In summary, the pre-opening investor sentiment of the stock was analyzed by BERT, and it was added to the LSTM model for predicting the closing price of the day, which

improved in all five evaluation indicators than the direct use of the LSTM model. Therefore, we believe that the addition of investor sentiment through BERT can effectively improve the accuracy of the LSTM model in predicting stock closing prices. In BERT-LSTM, the accuracy of predicting PingAn Bank's closing price is not improved much, and we think there are two reasons for this phenomenon. The first point is that the fine-tuned BERT is not accurate enough to analyze the investor sentiment of PingAn Bank. The second point is that PingAn Bank investors do not accurately judge the trend of PingAn Bank stock price before the stock opens, so the accuracy of predicting PingAn Bank stock price cannot be greatly improved after adding investor sentiment into the LSTM model.

5. Conclusions

In this paper, we propose a BERT-LSTM model for stock price prediction, which incorporates investor sentiment into stock prediction. The model is also used for PingAn Bank, ZTE and MuYuan in China A-shares and is compared with other time series forecasting models to demonstrate the effectiveness of the model in predicting stock prices. Meanwhile, the method proposed in this paper can provide investment advice to stock investors, which can be used to guide actual investment and has certain practical significance. However, due to the lack of time and the limited ability of the authors of this paper, this paper still has certain shortcomings and needs to be improved in future research. For example, the judgment of the sentiment classification dataset is not accurate enough and does not take into account the impact of stock market news and company financial reports on stock prices. In the future, we will combine other time series forecasting methods with the characteristics of the stock market to further study stock price prediction.

References

[1] B. Luo, Y. Chen and W. Jiang, "Stock Market Forecasting Algorithm Based on Improved Neural Network", Proceedings of 2016 Eighth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp. 628-631, 2016.

[2] Q. Yang and X. B. Cao, "Analysis and Prediction of Stock Price Based on ARMA-GARCH Model", Mathematics in Practice and Theory, Vol. 46, No. 6, pp. 80-86, 2016.

[3] A. A. Adebiyi, A. O. Adewumi and C. K. Ayo, "Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction", Journal of Applied Mathematics, Vol. 2014, No. 1, pp. 1-7, 2014.

[4] Z. Jin, Y. Yang and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM", Neural Computing and Applications, Vol. 32, No. 13, pp. 9713-9729, 2020.

[5] J. Xiao, X. Zhu, C. Huang, X. Yang, F. Wen and M. Zhong, "A New Approach for Stock Price Analysis and Prediction Based on SSA and SVM", International Journal of Information Technology & Decision Making, 2018.

[6] D. Zhang and S. Lou, "The application research of neural network and BP algorithm in stock price pattern classification and prediction", Future Generation Computer Systems, Vol. 115, pp. 872-879, 2021.

[7] H. Chung and K. S. Shin, "Genetic Algorithm-Optimized Long Short-Term Memory Network for Stock Market Prediction", Sustainability, Vol. 10, No. 10, 2018.

[8] W. Antweiler and M. Z. Frank, "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", Journal of Finance, Vol. 59, No. 3, pp. 1259-1294, 2004.

[9] J. A. Wurgler and M. P. Baker, "Investor Sentiment and the Cross-Section of Stock Returns", Economic Management Journal, Vol. 61, No. 4, pp. 1645-1680, 2006.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", Neural Computation, Vol. 9, No. 8, pp. 1735-1780, 1997.

[11] A. Graves, "Supervised Sequence Labelling with Recurrent Neural Networks", Studies in Computational Intelligence, Vol. 385, 2012.

[12] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, 2018.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need", Advances in neural information processing systems, Vol. 30 2017.

[14] X. P. Qiu, T. X. Sun, Y. G. Xu, Y. F. Shao, N. Dai, and X. J. Huang, "Pre-trained models for natural language processing: A survey", Science China Technological Sciences, Vol. 63, No. 10, pp. 1872-1897, 2020.

[15] H. Y. Zhou, S. H. Zhang, J. Q. Peng, S. Zhang, J. X. Li, H. Xiong and W. C. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting", Proceedings of AAAI, 2021.