Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 4, No. 3, 2023

EMD-BiLSTM Stock Price Trend Forecasting Model based on Investor Sentiment

Tianyu Xu, Xiaoling He

Zhejiang University of Science and Technology, Hangzhou 310012, China

Abstract: The movement of stock prices is the focus of investors' attention in the stock market, so stock price trend prediction has always been a hot topic in quantitative investment research. Traditional machine learning forecasting models are difficult to process nonlinear, high-frequency, high-noise stock price time series, which makes the prediction accuracy of stock price trends low. In order to improve the prediction accuracy, according to the temporal characteristics of stock price data, it is proposed to use a combination of empirical mode decomposition (EMD), investor sentiment and two-way long short-term memory neural network to predict the rise and fall of stock prices. Firstly, the empirical mode decomposition algorithm is used to extract the characteristics of the stock price time series on different time scales, and the investor sentiment indicators of the text from the close of the previous stock trading day to the opening of the next trading day are extracted by constructing a financial sentiment dictionary, and finally the EMD-BiLSTM model is used to predict the rise and fall of the next index trading day. Experiments on the dataset of stock price series show that the optimized BiLSTM model has strong predictive ability for the trend of consumer sector indexes.

Keywords: Bidirectional Long Short-term Memory Neural Network; Empirical Mode Decomposition; Investor Sentiment; Index Rise and Fall Forecasts.

1. Introduction

With the rise of cutting-edge technologies such as artificial intelligence and big data, machine learning methods have been widely applied to the research of stock price prediction and trading decisions [1], Zhang Xiao et al. [2] used Gradient Boosting Decision Tree (GBDT) to predict stock trends, and the results showed that the prediction effect of GBDT model was better than that of linear regression and random forest model. Shallow machine learning algorithms have simple structure, poor generalization ability, easy to fall into local optimal solutions [3], and are limited in processing raw format data. Deep learning learns effective feature representations from large amounts of input data by forming more abstract, high-level features through simple but nonlinear modules. The representative of deep learning, the Recurrent Neural Network (RNN), can consider the short-term correlation of time series, and the hidden layer receives not only the current data, but also the previous data information, so it can theoretically use data of any length of time; However, when RNNs learn sequences, they are prone to gradient vanishing and gradient explosion problems, resulting in their inability to grasp the nonlinear relationship over a long span. Based on this, Hochreiter [5] proposed a LongShort-Term Memory (LSTM) network, which solved the problem of RNN to a large extent through the gate mechanism. Chen et al. [6] used LSTM to predict the yield of the Chinese stock market and found that LSTM had the best prediction effect. Bidirectional Long Short-Term Memory (BiLSTM) is a combination of forward and backward LSTM layers, which can be used to extract two-way time series from the input sequence data, and theoretically can effectively capture the internal relationship of financial time series data. Yang et al. [7] used BiLSTM to predict the closing price of the CSI 300 Index, which verified the practical value of BiLSTM, a cutting-edge deep neural network, in the field of financial time series forecasting. Empirical Mode Decomposition (EMD) is an effective adaptive nonlinear time-varying signal information extraction method, which can decompose nonstationary nonlinear sequences into multiple Intrinsic Mode Function (IMF) and residual components with different frequencies and different physical meanings [8].

In order to help investors make better investment decisions and increase economic returns, this paper uses EMD method to extract new features of time series data, quantify investor sentiment, and use BiLSTM to learn time series features bidirectionally to improve the prediction effect. The consumer industry index was selected as experimental data for price prediction, and the accurate effectiveness of the proposed model was verified by comparing it with four typical stock price rise and fall prediction models, including LSTM, EMD-based long short-term memory network, SVM, and extreme gradient boost.

2. Introducing the Models

2.1. Empirical Mode Decomposition

EMD is an adaptive analysis method for processing nonstationary signals, which overcomes the defect that the basis function of wavelet decomposition needs to be set, and is suitable for processing nonlinear and non-stationary time series data.

In essence, EMD is a stationary process that separates the series into stationary fluctuation terms (IMF) of different scales and a residual trend term through a fixed pattern, and the time series to be decomposed has 3 assumptions:

1) The series to be decomposed has at least two extreme points, one maximum and one minimum.

2) The local time characteristics of the sequence to be decomposed are uniquely determined by the time scale between extreme points.

3) If the sequence to be decomposed has no extreme point and an inflection point, it can be differentiated one or more times to obtain the extreme value, and then the decomposition

result can be obtained by integration.

The specific decomposition method of EMD is as follows:

1) Determine all extreme points of the time series x(t) to be decomposed, use cubic spline interpolation method to fit the maximum and minimum, construct the upper envelope u(t) and the lower envelope l(t)

2) to calculate the mean m(t) of each time point of the upper and lower envelope, and the component h(t) is obtained by finding the difference between the time series x(t) and m(t) to be decomposed.

3) according to whether the new sequence $h_n(t)$ meets the IMF or stop criteria: if it does, it is retained as the IMF new component; Otherwise, x(t)= $h_n(t)$, n=n+1, and repeat the above steps until hn(t) reaches the standard.

4) Find the residual sequence:

$$r_i(t) = x(t) - imf_i$$

If $r_i(t)$ is not a constant value or monotonic function, the residual sequence $r_i(t)$ will be repeated again as a decomposition sequence Step 1) and Step 2). Finally, the original sequence is broken down into:

$$x(t) = \sum_{i=1}^{n} imf_i + r_n(t)$$

2.2. BiLSTM Neural Network

The LSTM network is an improvement on the topology of recurrent neural networks (RNNs). As a variant of RNN, the LSTM model effectively solves the gradient disappearance caused by the increase of network layers and the passage of time by introducing controllable self-loop, and its ingenious design structure is especially suitable for handling tasks with long delay and time interval. The structure of LSTM neural networks differs from other deep learning algorithms in that they have a special neuronal cell state, which can learn sequence data information that needs to be recorded and forgotten in a long-term state. Inside the cell unit, it consists of three gates: the forgetting gate, the input gate, and the output gate.

One is the forgetting gate, the first step in the operation of the LSTM model is to determine what interference information needs to be discarded in the cell state, then by reading h^{t-1},x^t in the sequence data, and giving different weights W_f and bias b_f between 0 and 1 to determine the degree of retention of sequence data information. The control function of the forgotten gate is as follows

$$f^{(t-1)} = \delta({}^{t}W_f[h^{(t-1)}, x^t] + b_f)$$

where b_f and W_f are biased and weighted for the forgetting gate, respectively.

The second is the output gate, the input gate controls the processing degree of the new input sequence data information added to the cell meta state, and two steps are required to achieve this process: first, the sigmoid function determines the degree of update of the information, as shown in Equation 2; Second, decide how much to update; Finally, the two parts are combined, the information that has been discarded in the oblivion gate is discarded and the updated information is added, stored in a long-term state

$$i^t = \sigma(W_i[h^{(t-1)}, x^t] + b_i)$$

$$c^t = \sigma(W_c[h^{(t-1)}, x^t] + b_c)$$

$$C^t = i^t * c^t + f^{(t)} * C^{t-1}$$

Where W_i and W_c represent the corresponding weights, b_i and b_c represent the corresponding biases, and C^t represents the current cell state value.

The third is the output gate, the LSTM structure will determine the output degree of sequence information based on the cell state, the output part of the cell state is judged by the sigmoid layer, and the long-term state of the cell element is processed with the tanh function and then multiplied with the previous output part, and finally the sequence information of the output is determined. The following is expressed by the formula

$$o^t = \sigma(W_o[h^{(t-1)}, x^t] + b_o)$$

$$h^t = o^t * \tanh^{-1}(c^t)$$

where W_o and b_o represent the weight and bias of the output gate, and h^t is the output value of the current cell.

BiLSTM is composed of two layers, forward LSTM and backward LSTM, in which the input time series is input into the LSTM model in the original order. In the backward LSTM layer, the input time series are entered into the LSTM model in reverse order. This structure extracts the bidirectional relationship of the time series and connects two layers of LSTMs to the same output layer. Therefore, the theoretical prediction performance should be better than that of unidirectional LSTM, and the specific expression of BiLSTM is as follows:

$$\overrightarrow{h_t} = \sigma\left(\overrightarrow{W_{xh}}x_t + \overrightarrow{W_{hh}}\overrightarrow{h_{t-1}} + \overrightarrow{b_h}\right)$$

$$\overleftarrow{h_t} = \sigma(\overleftarrow{W_{xh}}x_t + \overleftarrow{W_{hh}}\overleftarrow{h_{t-1}} + \overleftarrow{b_h})$$

$$H_t = \overrightarrow{W_{xh}}\overrightarrow{h_t} + \overleftarrow{W_{hy}}\overleftarrow{h} + b_y$$

where: σ is the activation function, and H_t is the hidden layer input. The final input is obtained by updating the forward and reverse structures. The structure of BiLSTM is shown in Figure 1.

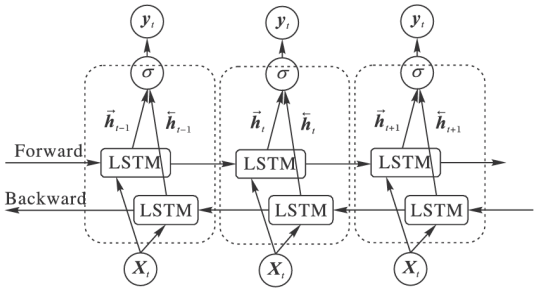

Figure 1. BiLSTM structure

3. Stock Price Rise and Fall Model Design

3.1. EMD Method Feature Extraction

In this paper, the EMD method is used for the initial time series, which is decomposed into local feature signals representing different time scales, and the EMD method is used to obtain its different time scale features for the stock closing price as new features as model input. In order to solve the problem of false high accuracy caused by EMD decomposition of data that needs to be predicted in the future, this paper adopts stepwise EMD decomposition for time series. First, the time series in the training set are decomposed by EMD, and then only the time series before the forecast are decomposed according to the forecast window, and the decomposition process is repeated until all data decomposition is completed, as shown in Figure 2. The EMD stepwise decomposition method can avoid the use of future data and improve the scientific and practical nature of the predictive model.

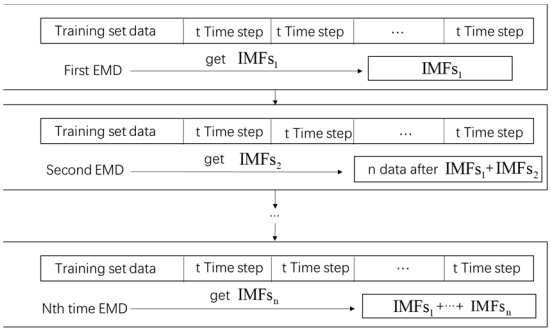

Figure 2. EMD stepwise decomposition method

3.2. Investor Sentiment Analysis

Use the investor sentiment to increase the model to input new features. Investor sentiment is to identify and summarize the emotional words in the text data, and then judge the investor emotions expressed in the text. At present, the methods of text analysis mainly include machine learning and sentiment analysis based on sentiment dictionary.

Sentiment analysis methods based on machine learning generally build classifiers based on supervised machine learning algorithms to achieve sentiment discrimination. This method has high classification accuracy, but it can only classify the sentiment of the text, and cannot quantify the emotional tendency into a specific value. The sentiment analysis based on the emotional dictionary corresponds the words in the emotional dictionary with the words after the text is segmented, obtains the negative and positive categories, and accumulates the words through the rules, and then obtains the emotional tendency score of the text. However, ordinary sentiment analysis based on emotional dictionaries does not consider the influence of modifiers in the text on emotional vocabulary, so this paper conducts sentiment analysis based on emotional dictionaries, and constructs auxiliary weight dictionaries when designing emotional dictionaries to optimize the weight of the basic weight of vocabulary.

At present, the three basic emotional dictionaries with the highest usage rate of sentiment analysis in Chinese are CNKI Emotional Dictionary, National Taiwan University Emotional Dictionary and Tsinghua University Emotional Dictionary. In this paper, these three dictionaries were merged, deduplicated and screened to obtain the initial basic emotional dictionary, including 7928 positive emotional words and 11785 negative emotional words, as shown in Table 1.

Table 1. Initial basic emotional dictionary contents

dictionary	vocabulary
Positive dictionary	Good news, beautiful scenery, happy, well-off, rich, wonderful, first-class, good, superior
Negative dictionaries	Disgrace, lack of morality, confusion, inferiority, deceived, bad

Table 2. Dictionary expansion of some vocabulary

dictionary	vocabulary
Positive dictionary	Soaring, must rise, bull stocks, bull market, takeoff, strong, bullish, up limit, high rich and handsome
Negative dictionaries	Plummeting, devaluation, make-up, big fall, drop limit, junk stocks, bear market, bird company, meat cutting

Texts in different fields have different text habits, unique vocabulary and network terms, and only using the initial basic dictionary for sentiment analysis of texts in the financial field cannot guarantee the emotional quality obtained, so the initial basic emotional dictionary is expanded with formal vocabulary and informal vocabulary in the financial field, of which 3183 positive emotional vocabulary and 2538 negative emotional vocabulary are expanded, as shown in Table 2.

The text dataset consists of investor exchange posts, stock information, and stock research reports. Investor communication posts are posts that investors share and communicate when trading stocks, and the content directly contains investor sentiment; Stock information includes information such as large stock transactions and company development; Stock research report is the research report of major institutions on stock companies and industries. Stock information and research reports affect investors' sentiment, thereby changing their investment behavior, which in turn affects the rise and fall of stocks. The sentiment analysis in this paper is based on the basic emotional dictionary and the auxiliary weight emotional dictionary, and the main process is as follows:

1) Carry out Chinese word segmentation of the text, and use word segmentation technology to divide the text into independent vocabulary units.

2) Match emotional words, compare the emotional words in the basic emotional dictionary, check whether the text contains emotional words, and retain the text if so; If not, the text is filtered.

3) Calculate the emotional score, optimize the weight of the basic weight of the basic emotional dictionary through the auxiliary weight emotional dictionary, and obtain the emotional score according to the definition formula, this document defines each positive word plus 1 point, each negative word minus 1 point, if there is an auxiliary weight emotional dictionary vocabulary before the vocabulary, the score is multiplied by the corresponding auxiliary weight. Obviously, the higher the score, the more positive the emotion; The smaller the score, the more negative the emotion.

Text specific sentiment score is calculated as follows:

$$score = \sum_{i=1}^{n} a * pos - \sum_{i=1}^{m} b * neg$$

where: score is the text sentiment score, a is the positive word auxiliary weight, $\sum_{i=1}^{n} a * pos$ is the positive word weight multiplied by the cumulative sum in the text, b is the positive word auxiliary weight, and $\sum_{i=1}^{m} b * neg$ is the negative word weight in the text multiplied by the accumulation and sum.

3.3. Overall Framework Design

In summary, the overall framework of the model consists of two parts, using EMD decomposition to obtain new characteristics for the closing price series, sentiment analysis to obtain daily investor sentiment, and finally input the obtained new characteristics, investor sentiment and basic stock market (opening price, closing price, high price, low price, trading volume, transaction amount) data into the BiLSTM neural network to predict the rise and fall of stock prices.

The hierarchy of EMD-BiLSTM neural network consists of 4 parts.

1) Input layer. Take the specific characteristics of each time step (including open, close, high,..., investor sentiment) as input. Let the input time step be n, then the input sequence

that predicts the rise and fall of the stock price on day t is expressed as follows:

$$X = \left[X_{t-n},\ X_{t-n+1}, \cdots,\ X_{t-1} \right]^T$$

$$X_{t-1} = \left[open_{t-1},\ close_{t-1}, \cdots, IMF_{t-1},\ Res_{t-1},\ sentiment_{t-1} \right]$$

Among them: IMF_{t-1} and Res_{t-1} are the values of the N IMF trend terms and residual terms obtained by decomposing the closing price on the t-1 day, and $sentiment_{t-1}$ is the investor sentiment on the t-1 day.

2) BiLSTM layer. Using two LSTM networks in the forward and reverse directions to learn the input sequence bidirectionally, using the input sequence X to calculate the output state $\overrightarrow{H_{t,i}}$ of the i^{th} input vector of the forward layer, and using the reverse form of X to calculate the output state $\overleftarrow{H_{t,i}}$ of the reverse layer, then the output of the BiLSTM layer is $H_t=[H_{t,1},H_{t,2},\cdots,H_{t,i},\cdots,H_{t,n}]^T$, where $H_{t,i}$ contains the forward $\overrightarrow{H_{t,i}}$ and the reverse $\overleftarrow{H_{t,i}}$:

$$H_{t,i} = [\overrightarrow{H_{t,i}} \oplus \overleftarrow{H_{t,i}}]$$

3) Dropout layer. Add a Dropout layer to prevent the model from overfitting. At each iteration, the output of BiLSTM neurons is randomly zeroed out according to the specified proportion.

4) Output layer. Select the fully connected layer to output the forecast value of the T-day closing price.

4. Empirical Analysis

4.1. Data Sources

The data used in this article comes from the Wind database, and the indicators include open, high, low, close, up or down, amplitude and turnover rate, etc., and the index is from January 4, 2010 to July 15, 2021. The text data such as investor communication posts, stock information, and research reports of the consumer index are from the Oriental Wealth Stock Bar Forum, covering the period from January 4, 2010 to July 15, 2021, with a total of 545598 articles, obtained using crawler technology. Among them, 85% of the data on the trading day is used as a training set to train the model, and the remaining 15% is used to test the model and verify the model effect.

4.2. Data Preprocessing

1) EMD decomposition

A stepwise EMD decomposition method is adopted for the closing price of stocks, and IMFs and trend terms are added as new features to the forecast model. Taking the current era as an example, the IMFs and trend items obtained after decomposition are shown in Figure 3

2) Text preprocessing

Text preprocessing includes the removal of duplicate text, the classification and summary of text from the close of the previous stock trading day to the opening of the stock market on the same day, word segmentation, de-pause words, and sentiment score quantification. Chinese word segmentation and removal of stop words: FoolNLTK based on BiLSTM training is used to complete word segmentation, and the stop word list of HIT after removing basic emotional vocabulary and auxiliary weight emotional vocabulary is used to complete the removal of stop words in the word segmentation process. Finally, sentiment quantification is completed according to the sentiment score calculation formula.

3) Normalized processing

In order to eliminate the influence of dimensional differences between input indicators and improve the speed and accuracy of model training, it is necessary to normalize the original data. This paper uses min-max standardization, and its calculation formula is as follows:

$$x^* = \frac{x - \min}{\max - \min}$$

where max is the maximum value of the sample data, min is the minimum value of the sample data, and x is the original value that needs to be normalized.

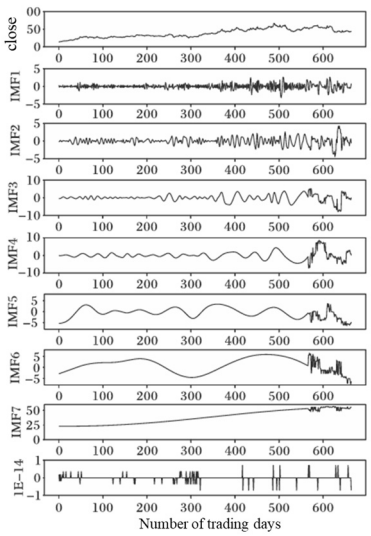

Figure 3. EMD decomposition results

4.3. Model Performance Comparison and Analysis

Table 3. Comparison of overall model indicators

model	accuracy	precision		recall		F1	
		Rise	Fall	Rise	Fall	Rise	Fall
BiLSTM	56.51	58.71	58.78	58.95	58.05	58.73	58.60
EMD+ BiLSTM	61.54	60.36	60.50	62.45	61.99	61.80	61.07
emotion +BiLSTM	62.33	63.12	65.16	62.60	62.10	63.31	62.53
EMD+emotion+ BiLSTM	67.12	68.41	68.32	63.21	68.13	64.44	68.53

In order to evaluate the model proposed in this paper, this paper conducts a series of comparative experiments and sets the time window of the model input series to 10, that is, based on the data of the previous 10 stock trading days (10 is the time step), predict the rise and fall of the stock price in the next stock trading day in the future. Taking the basic market data of stocks (opening price, closing price, low price, high price, total trading volume, and total trading volume) of each stock trading day as the basis, the IMFs, trend items and investor sentiment obtained by EMD decomposition were

added respectively, and the EMD-BiLSTM model was established, and the comparison experiment with BiLSTM was carried out. Table 3 shows the comparison of the overall prediction indicators of each model in the consumer industry index.

It can be seen from Table 3 that the introduction of affective factors and EMD in the model gradually improves the prediction accuracy of the model, which can enable the model to better learn more important information. The overall prediction effect of BiLSTM based on empirical mode decomposition and investor sentiment proposed in this paper is better than other models, and the prediction performance of rise and fall is more balanced than that of other models.

Table 4. Comparison of common model indicators (input maximum optimization)

model	accuracy	precision		recall		F1	
		Rise	Fall	Rise	Fall	Rise	Fall
EMD+BiLSTM	71.91	70.66	71.75	70.14	69.19	72.34	69.20
BiLSTM	64.09	63.52	63.78	63.00	63.99	64.20	63.07
EMD+LSTM	61.36	60.80	60.62	60.34	61.31	62.76	62.86
LSTM	59.93	61.26	60.51	57.17	58.50	58.31	60.99
SVM	57.29	56.71	56.68	57.50	57.09	56.23	56.04
XGBoost	53.18	55.25	55.38	52.89	50.36	51.33	55.39

In addition, in order to compare with more common stock price rise and fall prediction models, this paper adds EMD-LSTM, LSTM, SVM, XGBoost models to compare consumer industry indices under the condition of maintaining maximum input optimization (that is, all model inputs include stock market data, EMD, and investor sentiment). It can be seen from Table 4 that BiLSTM can perform two-way time series feature learning, which can better learn the stock market trend than LSTM, and EMD decomposes the index, making the data information presentation easier to be understood by the model and improving the prediction effect. The EMD-BiLSTM model outperforms other models in all indicators. In summary, it is proved that the model proposed in this paper has greatly improved the accuracy of stock price rise and fall prediction.

5. Conclusion

Aiming at the problem of large prediction error caused by nonlinearity, non-stationarity and chaos of stock time series, this paper proposes a BiLSTM stock price rise and fall model based on empirical mode decomposition and investor sentiment. This paper uses EMD to decompose the closing price to obtain the new characteristics and trend terms of the closing price on different time scales, thereby obtaining the new features. On the basis of the initial basic emotional dictionary, the formal vocabulary and informal network vocabulary in the financial field are expanded, and the auxiliary weight emotional dictionary is constructed to optimize the weight of the basic weight of the expanded basic emotional dictionary, and the emotional score corresponding to the text is calculated. These new features are added to BiLSTM for bidirectional time series feature learning. Through experimental comparison, it can be seen that the prediction model can better predict the stock market trend through EMD and investor sentiment, which has a certain guiding effect on investors. At present, most stock forecasting research only uses the basic indicators of the stock market, and ignores the impact of investors' investment decisions and behaviors on the stock market, so the prediction effect is average. The model proposed in this paper extracts in-depth information from the basic market indicators of the stock market, conducts in-depth analysis of investor sentiment that may change investors' investment decisions, performs two-way learning on time series characteristics, and selectively pays attention to the influence of different components on the prediction results, so that the model prediction has better effects. In future work, more factors can be added, such as stock market announcements, policies, and other technical indicators, to further improve the accuracy of model prediction.

References

[1] Zhang X, Zhang Y J, Wang S Z,et al. Improving stock market prediction via heterogeneous information fusion[J]. Knowledge-Based Systems,2018,143:236-247.

[2] ZHANG Xiao, WEI Zengxin, YANG Tianshan. Application of GBDT Combinatorial Model in Stock Prediction[J]. Journal of Hainan Normal University(Natural Science Edition), 2018, 31(1):73-80.

[3] WANG S, WANG X, WANG S, et al. Bi-directional long short term memory method based on attention mechanism and rolling update for short-term load forecasting [J]. International Journal of Electrical Power and Energy Systems, 2019, 109: 470-479.

[4] Kexin H, Zhijin Z. Prediction stock price based on CNN and LSTM models [J]. Financial Engineering and Risk Management, 2022,5(7).

[5] HOCHREITER S. Untersuchungen zu dynamischen neuronalen netzen [D]. Munich: Technische Universitat Munchen, 1991:91.

[6] CHEN K, ZHOU Y, DAI F. A LSTM-based method for stock returns prediction: a case study of China stock market[C]// Proceedings of the 2015 IEEE International Conference on Big Data. Piscataway: IEEE, 2015: 2823-2824.

[7] YANG M, WANG J. Adaptability of financial Time Series Prediction Based on BiLSTM [J].Procedia Computer Science, 2022,199: 18-25.

[8] WENG Xiaojian,LIN Xudong,ZHAO Shuaibin. Long-term short-term memory network stock price rise and fall prediction model based on empirical mode decomposition and investor sentiment [J]. Computer applications,2022, 42（S2）: 296-301.