DOI: 10.53469/jgebf.2022.04(04).12

# Predictive Analysis of Stock Price by Machine Learning and Online Investors' Sentiment

**Yanguo Li**

School of Finance, Yunnan University of Finance and Economics, Kunming, China

**Abstract:** ***Investors in the market are not absolutely rational, and investors' sentiment will have an impact on investment behavior, which will eventually extend to the entire market. In this paper, based on the data from Baidu Index, the online investors' sentiment index was constructed. Then, by using BP neural network, two prediction models with and without investors' sentiment index were constructed respectively to predict the stock market data, and compared the prediction results with or without the sentiment index. Many conclusions can be drawn from research. The BP neural network model has high prediction accuracy and can effectively predict the average price of the Shanghai Stock Exchange. After adding the online investors' sentiment index into the neural network, the prediction accuracy of the model is significantly improved, which shows that the investors' sentiment index based on the network data information can better reflect the investors' sentiment, and more comprehensive market information can make the forecasting model better learn the relationship between market data, so as to improve the accuracy of prediction. Investors' sentiment has a positive impact on the prediction of China's stock price. The addition of sentiment index can make the prediction effect better.***

**Keywords:** Investors' sentiment, Baidu Index, Neural network, Stock price prediction.

## 1. Introduction

The stock market plays an important role in the prosperity and development of the economy. In the actual investment process, investors are susceptible to emotional information when facing many stocks. After the sentiment of a single investor is spread, group sentiment will gradually form, and the influence of individual decision-making will be amplified, which will eventually cause the fluctuation of the entire stock market.

With the development of Internet information technology, a large number of Internet users search for investment information, express opinions and express emotion through Internet platforms. This "wisdom of the crowd" provides investor sentiment that can represent market sentiment. Baidu can provide a large amount of timely data and information. More real investor sentiment information can be obtained from the data information, which is of great significance for research. In actual financial activities, investment institutions can obtain required information from the data from Baidu to construct sentiment indices to achieve the purpose of predicting investment behavior, and use it in the investment decision-making process to improve the efficiency of investment decision-making. As for the financial supervision department, timely grabbing stock market information from Baidu's database can form an efficient public opinion monitoring mechanism. In turn, the supervisory system can be better improved and the stability of the stock market can be maintained.

Artificial intelligence is also a hot topic. As the core of artificial intelligence, machine learning has also made great progress, and this method can also fully simulate the laws of the stock market. The fluctuation of stock price directly affects the stability of the stock market and the healthy development of the financial market and the national economy. A successful stock price forecast can help investors make profits, and it can also help government departments to provide timely and reasonable market guidance and supervision. As research on machine learning becomes more complete, it becomes a better choice to apply machine learning to stock market prediction.

## 2. Literature Review

From the perspective of psychology, it can be found that in addition to information, the role of emotion in human decision-making is also very prominent. In financial markets, public sentiment is equally powerful. As the Internet has become an indispensable part of people's lives, more and more people use the Internet to access information and express their emotions. The information on the Internet can be understood as a form of collective intelligence and can reflect the user's emotions, which has an excellent predictive effect on some information in the real world. Gruhl et al. (2005) use the Internet as the basis of the research environment, select chat information related to a book, and use it as an independent variable to predict the future sales of the book. When the blog was still popular, people often posted their thoughts on the blog. For example, people's preferences for a movie can be expressed through blog, and the future sales of the movie can be predicted based on this emotional state (Mishne et al., 2006). Huberman (2010) sums up the corresponding public sentiment based on the opinions of a movie on the Twitter platform, so as to predict the future box office of the movie. In the research, Hyunyoung (2011) predicts car sales, the number of tourists in a tourist destination and so on based on Google's search volume for a certain keyword. The huge amount of data and information that can be updated in real-time on the Internet has great advantages in prediction.

Using information from the Internet in the stock market can also produce excellent forecasting results. Bollen et al. (2011) take the information related to the stock market on the Twitter platform as the object of the research, so as to create the corresponding calm sentiment indicator, and then use it to predict the Dow Jones Industrial Average. The effect is very ideal. Karabulut (2013) creates the GNP index by selecting the information about securities on the Facebook platform in

the research, so as to predict the stock market. In fact, the emotions reflected by the information on the Internet will have a significant impact on the final investment decision, and then affect the stock market.

The stock market is inherently dynamic, nonlinear and complex, so using traditional linear regression methods is not the best choice. Using machine learning methods can get more accurate prediction results. With the development of machine learning theories, many scholars study the application of machine learning models in the stock market. H.White (1988) takes IBM's daily rate of return as a prediction object to conduct research, and uses neural network to predict it, but the model of this experiment doesn't meet the expectation. Gencay (1996) successfully creates a forward neural network model, and carries out an empirical study on the stock price data of the Dow Jones Industrial Average, and obtains good prediction results. Kinjal et al. (2011) take the Indian stock market as the research object, predict through the 3-layer BP neural network, and finally obtain good prediction effect by adjusting the number of input and hidden layers in the process. Persio et al. (2017) use the opening price, closing price, highest price, lowest price and trading volume as input data to compare RNN, LSTM network and Gru for predicting future stock price. The results show that LSTM network performs best. Zhang et al. (2017) extract investor sentiment and transaction data from social media networks and combine LSTM network to predict the rise and fall of stock price. The experimental results show that the prediction performance of LSTM network can be effectively improved. Nelson et al. (2017) compare the performance of LSTM network, SMLP and RF for stock price prediction. The results show that LSTM network performs best. Hoseinzade et al. (2019) design two kinds of convolutional neural networks to predict ups and downs of stock index, both of which are superior to the benchmark model. Yu et al. (2020) use phase space reconstruction method combined with LSTM network for stock price prediction. The experimental results show that the model has better prediction accuracy. From these research results, the application of this technology to stock market prediction has excellent performance.

Because of the advent of the big data era on the Internet, many scholars analyze investor sentiment based on the huge and high-frequency data in the network, and then build sentiment indices. At the same time, with the rise of machine learning, it has become a new idea to extract sentiment from the network and use this method to study its impact on stock price prediction.

## 3. Methodology and Date

### 3.1 Neural Network

The neural network is an adaptive method for self-training on the basis of original data. The advantage of neural networks in the field of prediction lies in their self-learning ability and self-adaptive ability. Even when the input-output relationship of the entire neural network system is very complex, it can well find the internal relationship between the input and the output. Therefore, in the case that it is difficult to find out the relationship between the data, the problem of complex relationship between the data can be well solved after scientific and reasonable training of the data set using the neural network.

Neural networks can recognize new patterns after training, even if the training set doesn't contain these new data. Predicting future data is based on past data, so neural networks have great advantages in prediction problems. In addition, neural networks have strong function approximation capabilities. A three-layer neural network can be used to approximate any complex nonlinear function.

Neural networks have very powerful learning capabilities. The behavioral finance theory believes that the dynamic changes of data in the stock market are the result of people's psychological changes, so investment sentiment can be reflected in the dynamic changes of the stock market to a large extent. Using neural networks to study the impact of investors' sentiment on stock price prediction is a good choice.

### 3.2 Data and Index Construction

3.2.1 Date

Baidu provides great convenience for stock market investors to obtain stock market information. Users' opinions and emotion can be expressed here, and these behaviors can also reflect investors' emotional state in essence. This is also the logical basis for the research on using the data on the Internet to predict the stock market.

Investors can freely search for the information they need by using Baidu, which naturally shows the most real psychological emotion in this case. Therefore, the information obtained from Baidu is also closer to the real sentiment of investors. Baidu generates a lot of data all the time. Compared with traditional sentiment indicators, the online sentiment data can reflect users' mood and behavior changes in real-time and more accurately, which provides very excellent data source for academic research.

This paper selected the Baidu Index between January 4th, 2016 and September 30th, 2021, and used daily data in terms of data frequency. The data from Baidu Index of the keywords being "bull market" and "bear market" was selected, and the data corresponding to each trading day of the stock market was filtered out.

SSE Composite Index is an important index that reflects the overall volatility of all stock prices on the Shanghai Stock Exchange. This paper used data from this index to not only reduce the impact of individual stocks, but also represent the entire stock market. This paper selected the daily data of the opening price, closing price, highest price and lowest price between January 4th, 2016 and September 30th, 2021, and then calculated the average of the four values as the average price of one day.

3.2.2 Index construction

The investors' sentiment index in this paper drew lessons from the practice of Mao, Scott and Johan (2015), and defined the investors' sentiment index as:

$$ISI_{baidu} = In(1+N) - In(1+X) \quad (1)$$

Among them, *N* and *X* represented the Baidu Index daily data when the keyword was "bull market" and "bear market" respectively.

## 4. Empirical Analysis of Stock Price Prediction

### 4.1 Design of Neural Network Model

4.1.1 Network structure

It's the core of determining its network structure that the BP neural networks used in this paper determine the number of hidden layers and the number of hidden layer unit nodes. Through experiments, it could be found that compared with increasing the number of hidden layers, we increased the number of neurons in the hidden layers to reduce the error, and the final training effect was better achieved. Under the condition of reasonable structure and appropriate weights, a 3-layer BP neural network can approximate any continuous function. Therefore, the hidden layer number was set to 1 in this paper to obtain a 3-layer BP neural network.

The performance of neural networks is closely related to the number of nodes in the hidden layers. For the determination of the number of hidden layer nodes, the initial value is generally obtained based on the number of input variables and output variables and previous experience, and then the number of hidden layer nodes is gradually increased or decreased. Finally, the number of hidden layer neurons with the best effect was selected through experiments. It could be determined that the number of neurons in the hidden layer was nine.

4.1.2 Main parameters

The selection of learning rates, maximum training times, weights and thresholds is very important for BP neural networks. In order to ensure the stable operation of the neural network model, the value of the learning rate is generally between 0.01 and 0.8, and then the error drop curve is observed in the experiment, so as to select the learning rate that can make the error curve drop faster and the experimental results be the most satisfactory. Finally, the learning rate in this paper was set to 0.01. For the number of learning times, if it is blindly increased, it will only lead to oscillation or divergence of the neural network, so as to reduce the success rate of network training. Through repeated debugging, the upper limit of training times was finally set to 5000 times. The initial weights and thresholds are of great significance to the models, which can not only reduce the error of the network prediction, but also avoid the occurrence of pauses in the training process. Based on the original input data and output data, the weights and thresholds were optimized through the training process, and the appropriate values were finally determined.

### 4.2 Prediction of Stock Market Based on Neural Network

In this paper, BP neural network was used to predict the stock market, and two prediction models were constructed. The first model predicted the corresponding value of the next trading day (t) through the average value of the SSE Composite Index in the previous 4 stock market trading days (t-4 to t-1). The second model was to add the investors' sentiment index of the previous 2 trading days (t-2 and t-1) to predict the corresponding stock index of the next trading day (t) on the basis of the average value of the SSE Composite Index in the previous 4 trading days (t-4 to t-1).

4.2.1 Prediction results of the neural network model without sentiment index

From Figure 1, it can be roughly seen that the prediction model has a relatively good forecasting effect on the average value of the SSE Composite Index. The predicted value of the model was roughly the same as the overall trend of the actual value of the stock index, which indicated that the predicted value obtained by using the data of the previous 4 trading days to predict the average price of the next trading day had a high degree of fit with the corresponding actual value. The overall prediction accuracy was relatively high.

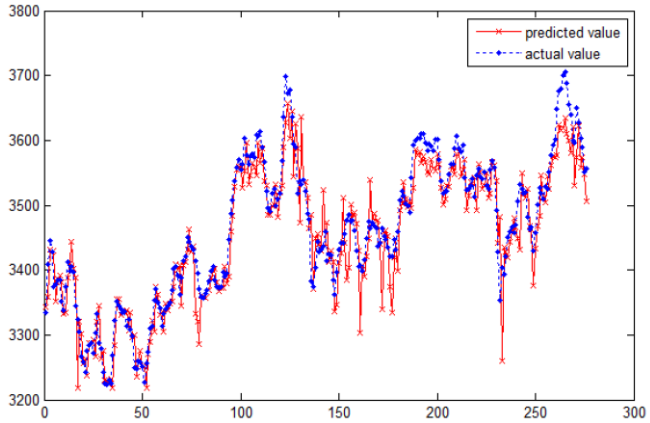


**Figure 1:** Prediction results without sentiment index

4.2.2 Predictability analysis

This paper need analyze the predictability of the forecasting model according to the accuracy of stock market data forecasting. Therefore, this paper conducted a specific analysis by constructing a predictability index. The calculation formula of the index is:

$$P = 1 - \left| \frac{y_p - y_a}{y_a} \right| \quad (2)$$

$y_p$ is the predicted value and $y_a$ is the actual value. This index can reflect the distance between the absolute percentage of the prediction error to the true value and 100%. This index can directly reflect the accuracy of all individual prediction results, thus reflecting the prediction effectiveness of the BP neural network. The smaller the predictability index is, the larger the difference between the predicted value and the actual value is; conversely, the larger the predictability index is, the closer it is to 1, the closer the predicted value is to the actual value. The predictability index is 1 when the predicted value is equal to the actual value. The index can well reflect the similarity of each predicted value to the actual value, so as to obtain the predictability of the prediction model for stock market data.

As can be seen from Figure 2, the predictability indices are all greater than 0.95, and the predicted values are relatively close to the actual values, without particularly large deviation. From

this, it can be seen that the overall accuracy of using the neural network to predict the stock market is relatively high, and the predicted values are generally close to the actual values. In predicting the stock market, this method of neural network has great advantages and is an effective method.

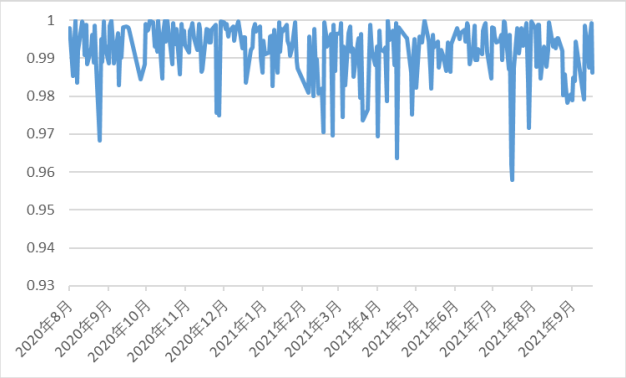


**Figure 2:** Time series diagram of predictability index

4.2.3 Prediction results of neural network model when adding sentiment index

The prediction effect is shown in Figure 3. After adding the investors' sentiment index based on Baidu Index, the prediction effect of the prediction model for the average price is significantly better than that of the model without adding the online investors' sentiment index. After adding the investors' sentiment index based on Baidu Index, the prediction accuracy of the neural network model for predicting stock market data was significantly improved.

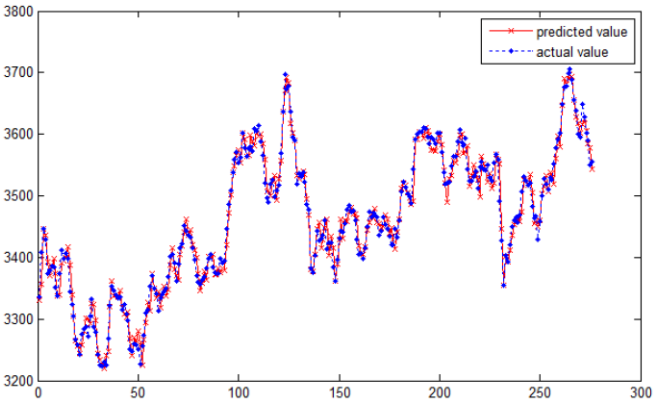


**Figure 3:** Predicted results with sentiment index

### 4.3 Comparison of Prediction Accuracy of Different Models

4.3.1 Evaluation Indices of model prediction

In this paper, mean squared error (MSE) and mean absolute percentage error (MAPE) were used to evaluate and compare the prediction ability of the prediction models. The smaller the values of these two indicators are, the closer the predicted values are to the actual values, and the better the prediction effect of the neural networks is.

Mean squared error and mean absolute percentage error are defined as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) \quad (3)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \quad (4)$$

$N$ is the number of samples, $y_i$ is the ith actual value, and $\hat{y}_i$ is the ith predicted value.

4.3.2 Comparison and analysis of evaluation indices of various models

The evaluation indexes of the prediction ability of BP neural network models to predict the average price are shown in Table 1. It can be seen from this table that whether it is the mean squared error or the mean absolute percentage error, the error indexes values obtained from the prediction after adding the online investors' sentiment index based on Baidu Index are smaller than the error indexes values of the model without adding the index. It shows that the prediction accuracy of the neural network models is significantly improved after adding the online investors' sentiment index, and the addition of the investors' sentiment index can improve the prediction ability of the prediction model for China's stock market.

The experimental results are in line with the characteristics of China's stock market. The ups and downs of China's stock market are greatly affected by policies and positive and negative news. Moreover, there are many ordinary retail investors who lack investment experience and professional knowledge in China's stock market. They are more vulnerable to the influence of emotion in the market and become irrational. Investors' sentiment has a great impact on the stock market. Obtaining investors' sentiment through the huge data and information in the network can better reflect the investment sentiment of stock market investors, which is beneficial to our research and analysis in this area.

**Table 1:** Evaluation standards of forecasting performance

| | no sentiment index | sentiment index |
|---|---|---|
| mape | 0.797% | 0.351% |
| mse | 1441.712 | 262.369 |

## 5. Conclusions

Because Baidu Index can more intuitively reflect the behavior and sentiment of investors on the Internet, and the data is huge and easy to obtain, this paper used Baidu index to reflect investors' sentiment. Because the neural network can obtain the ideal forecasting effect in the complex and changeable stock market, this paper constructed the BP neural network forecasting models. This paper used the models with and without the investors' sentiment index to predict the stock market data, and then compared the model prediction results.

The following conclusions can be drawn from the research in this paper. The neural network models fitted the average price well. Because of its strong learning adaptability, nonlinear mapping ability and function approximation ability, the neural network can solve many problems that cannot be solved by traditional measurement methods. In this paper, the neural network models were used for prediction and satisfactory results were obtained. This method is very effective in predicting stock price. After adding the online investors' sentiment index to the neural network model, the prediction accuracy of the model was significantly improved. First, it can be shown that investors' sentiment in China has a

non-negligible impact on the stock market, and the investors' sentiment index based on network data information can better reflect the emotional state and changes of stock market investors. Second, it can be explained that having more comprehensive market information data can make the forecasting model better learn the relationship between market data, thereby improving the accuracy of forecasting. Finally, it can be shown that the addition of investors' sentiment does have a positive impact on the prediction effect of China's stock price, which can make the prediction results more accurate. The addition of online investors' sentiment can improve the predictability of stock price.

## References


[1] Antoniou C, Doukas J A, Subrahmanyam A. (2015). Investor sentiment, beta, and the cost of equity capital. Management Science, 62(2), 347-367.

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

[3] Chong O, Sheng O. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. International Conference on Information Systems, Icis 2011, Shanghai, China, December. DBLP.

[4] Firth M, Wang K P, Wong S M L. (2015). Corporate transparency and the impact of investor sentiment on stock prices. Management Science, 61(7), 1630-1647.

[5] Gencay. R. (1996). Non-linear prediction of security returns with moving average rules. Journal of Forecasting, 15(3), 43-46.

[6] H. White. (1988). Economic prediction using neural networks: the case of IBM daily stock returns. Neural Networks, IEEE International Conference, 2(6), 451-458.

[7] Hoseinzade E, Haratizadeh S. (2019). CNNpred: CNN-based stock market prediction using a diverse set of variables. Expert Systems with Applications, 129, 273-285.

[8] Huang D, Jiang F, Tu J, et al. (2014). Investor sentiment aligned: A powerful predictor of stock returns. Review of Financial Studies, 28(3), 791-837.

[9] Huberman. (2010). Predicting the future with social media. Social Science Electronic Publishing, 7(2), 492-499.

[10] Karabulut Y. (2013). Can facebook predict stock market activity? Social Science Electronic Publishing.

[11] Li, J., Chen, Y., Shen, Y., Wang, J., Huang, Z. (2019). Measuring China's stock market sentiment. Available at SSRN 3377684.

[12] Mao H, Counts S, Bollen J. (2015). Quantifying the effects of online bullishness on international financial markets. Statistics Paper.

[13] Nelson D M Q, Pereira A C M, Oliveira R A D. (2017). Stock market's price movement prediction with LSTM neural networks. International Joint Conference on Neural Networks, 1419-1426.

[14] Persio L D, Honchar O. (2017). Recurrent neural networks approach to the financial forecast of Google assets. International Journal of Mathematics and Computers in simulation, 11, 7-13.

[15] Pyo, D-J. (2017). Can big data help predict financial market dynamics? Evidence from the Korean stock market. East Asian Economic Review, 21(2), 147-165.

[16] Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. Journal of Banking & Finance, 84, 25-40.

[17] Schmeling, M. (2009). Investor sentiment and stock returns: Some international evidence. Journal of Empirical Finance, 16(3), 394-408.

[18] Yu P, Yan X. (2020). Stock price prediction based on deep neural networks. Neural Computing and Applications, 32, 1609-1628.

[19] Zhang G, Xu L, Xue Y. (2017). Model and forecast stock market behavior integrating investor sentiment analysis and transaction data. Cluster Computing, 20(1), 1-15.

[20] Zhang, X., Shi, J., Wang, D., Fang, B. (2018). Exploiting investors social network for stock prediction in China's market. Journal of Computational Science, 28, 294-303.