Based on Baidu Index and GBDT Shanghai Index rise and fall forecast

BoWen Xiu*

School of Mathematics and Statistics, ChongQing Technology and Business University
ChongQing, China

*Corresponding author: 2018104110@email.ctbu.edu.cn

Abstract. The stock market reflects the country's economic conditions, and it is of great significance to have a good prediction effect on the stock market. But with the rapid rise of the Internet, big data, and machine learning, the prediction of the stock market trend is not limited to the traditional methods and data sets. The trend of the stock market is not only dependent on itself but also affected by some other factors. Therefore, based on the machine learning model, this paper studies the prediction of investors' attention to the Shanghai Composite Index trend. This paper crawled the relevant index data from the website of Baidu Index based on the selected keywords. The correlation coefficient is used to select the keyword data with the best lag order and data type and as the model's input data. Through the establishment of LSTM, LASSO, RF, and GBDT models, the rise and fall of the Shanghai Composite Index are predicted. That is to say. The paper takes the accuracy of the rise and fall prediction as the judgment standard. GBDT model has the best prediction effect on the Shanghai Stock Exchange Index and can best explain the rise and fall of the Shanghai Stock Exchange Index. So, people can use this research to buy stocks before they rise and sell them before they fall.

Keywords: The stock market; Baidu index; Machine learning.

1. Introduction

The stock market reflects the condition of a country's economy, and its normal and healthy operation can promote the country's long-term economic development. Although the development of China's stock market is not mature, and its development momentum is strong. In the investment of the real economy, the financing from the stock market accounts for a considerable part, so it has a certain supporting role for the development of the real economy.

Nowadays, time series forecasting, grey model forecasting, and machine learning methods have been used to predict the stock market. Agarwal et al. had made good predictions on the Indian stock market based on the traditional time series models such as AR, MA, and ARMA [1]. Angadi et al. made a relatively accurate prediction of the stock market based on the ARIMA time series model [2]. Sun used experiments to show that machine learning can predict results better than traditional statistical models [3]. Chen et al. used the stochastic prediction model and the LSTM model to forecast the Chinese stock market. The results proved that the prediction effect under the LSTM model was significantly better than that of the stochastic prediction model [4]. But it turns out that similar predictions using similar indicators tend to get worse over time. At the same time, it believes that the ups and downs of the stock market are related to the data in the stock market and have a great relationship with the sentiment and attention of investors.

In 2006, after Google published search data queries for any keyword, people realized that the attention that could be expressed through search could have a huge impact on some industries. Choi et al. used the Google index to successfully predict car sales, unemployment claims, travel destination planning, and consumer confidence [5]. Kissan Joseph et al. and Bijl L et al. found that the Google index could be used to predict the stock market data in the short term with great accuracy [6, 7].

Baidu index is similar to Google index, which is based on Baidu web search and Baidu news search and reflects the attention of netizens to a certain issue. Therefore, through the analysis of the Baidu index, it can quickly discover a hot event or news on the network and quickly obtain the attention of netizens to an event. Therefore, based on this idea, the search of the Baidu index related to the stock

market can also obtain the netizens' concerns about the stock market, so the use of the Baidu index to predict the stock market has a great reference value. The study of Wang shows that high attention often leads to a high return rate, which confirms the correlation between investors' attention and stock market fluctuations [8].

In general, most of the research on the Baidu index is focused on selecting a stock as a keyword to query the Baidu index and make predictions.

2. Selection of data

To forecast the stock index, the data of the Shanghai Composite Index are divided into two categories. One is the data in the market. The other is the data outside the market. Tushare is a free, open source Python financial data interface package. The main implementation of the stock and other financial data from data collection, cleaning, and processing to the process of data storage can provide financial analysts with fast, clean. A variety of easy analyze the data, for them in the aspect of data acquisition to greatly reduce the workload, so they are more focused on the strategy and model research and implementation. Given the advantages of the Python Pandas package for quantitative financial analysis, most of the data returned by Tushare are in the Pandas DataFrame type, making it easy to use Pandas/Numpy /Matplotlib for analysis and visualization. Therefore, the data in the market are obtained by obtaining the API provided by the Tushare platform and using the web crawler to obtain the corresponding data. The selected data are the daily data of the Shanghai Composite Index from January 4, 2016, to December 30, 2020.

The data outside the market belongs to the Baidu index under certain keywords, representing the Internet search. Under the relevant keywords, the greater said Internet users search volume is larger. Research chooses the day and date of stock data corresponding to the degree of data, due to the rapid development of modern information technology and mobile devices, research divides Baidu indexes under keywords into PC, Mobile, and PC + Mobile three categories. To avoid the particularity of choice, research has collected more than 30 keywords such as "plate", "volume", "hit new", "loan", "dark horse", "profit", etc. Among which these keywords are not only related to the professional vocabulary of the stock market but also some of the "saliva words" that investors often say. Previous studies often rarely choose some "verbal" words that stock investors often speak in the stock market, so the sample selected in this paper is more comprehensive and representative.

Considering the problem of time delay, this paper will compare the correlation degree between keywords and the stock market under different time delays. The research divides the data collected under keywords into three categories. The first category is pre-time keywords. That is, the trend of these keywords will have a leading trend for the trend of the Shanghai Composite Index. The second category is simultaneous keywords. That is to say, this kind of keywords and the Shanghai Composite Index change the same trend; The third category is time-delayed keywords, that is to say, the trend of these keywords lags behind the trend of the Shanghai Composite Index.

In order to predict the accuracy and generality of the experimental results, the explanatory variables of the model should be selected with the highest correlation. For example, under the keyword "plate", it should select the lagged second order and its Baidu index at the PC end. Therefore, a maximum correlation coefficient Z is defined here:

$$\mathrm{Z}=\max\{x_{ij}\},(i=1,2,3;\ j=1,2,3,4) \tag{1}$$

x_{ij} is the value of the corresponding position in the table, x_{11} is the value of the first row and the first column.

Tab 1. Correlation Coefficient of Keywords.

keywords	Z	keywords	Z
Plate	0.049437	Value investing	-0.088018
Trample	-0.038048	leek	-0.093230
Volume	0.168817	Financial management	0.066794
Chip	-0.034066	Tap	0.038224
Subscription for new shares	0.095227	Tourism	0.065120
Loan	-0.009777	Enterprise	-0.096274
Stock is bullish	-0.042225	Trend	-0.060450
Rebound	0.021430	Daily necessities	0.025559
Concept	0.036381	Retail	-0.077425
Liquidated at a loss	-0.305781	Business	0.036113
Industrial	0.047224	Hold up	-0.212744
Shopping	-0.160600	Strong point	-0.045112
Investment loss	-0.532447	Banker	-0.075101
Advertising	-0.038655	Rent	-0.019796
Dark horse	-0.031083	Reaching peak	-0.028410
Profit	-0.049177		

Due to the limitation of the value of the correlation coefficient obtained, the keywords with the maximum absolute value of the correlation coefficient above 0.05 were selected as the final keywords. The selected keywords are shown in the table below:

Tab 2. Highly Relevant Keywords.

keywords	Z
Volume	0.168817
New play	0.095227
Liquidated at a loss	-0.305781
Shopping	-0.160600
Investment loss	-0.532447
Value investing	-0.088018
Leek	-0.093230
Tourism	0.065120
Retail	-0.077425
Hold up	-0.212744
Banker	-0.075101

3. Introduction to machine learning models

3.1 LSTM:

The general RNN model has a weak ability to describe the time series data with a long memory. When the time series is too long, it becomes very difficult to train RNN because of the phenomenon of gradient dissipation and gradient explosion. The LSTM model was modified based on RNN structure so that the problem that the RNN model could not describe the long memory of time series was solved. It solves the phenomenon of gradient disappearance and gradient explosion in the process of backpropagation. Introducing the gate's mechanism solves the problem of long memory that the RNN model does not have. The structure diagram of the LSTM model is shown in the following figure:

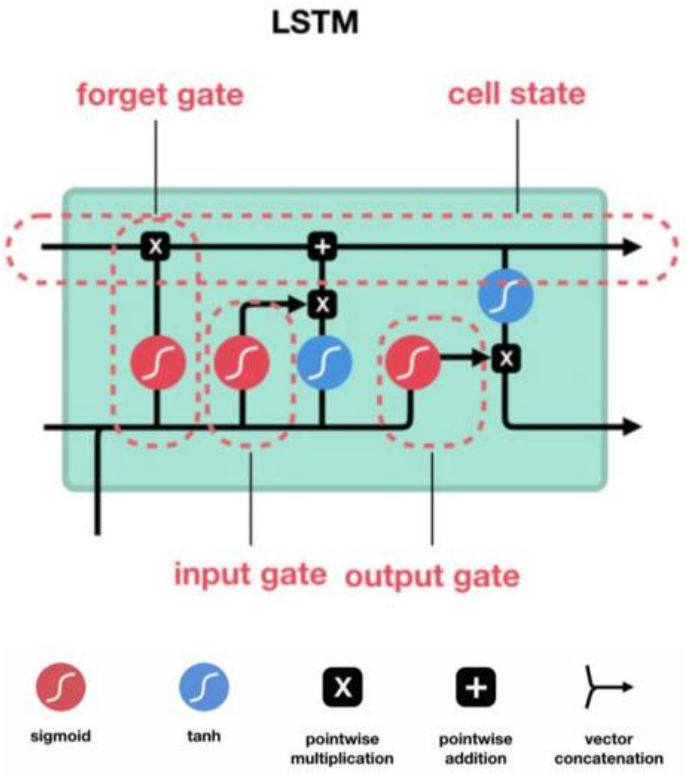

Fig 1. LSTM flowchart.

The core concept of LSTM is the cell state and "gate" structure. Cellular states act as pathways for information to be passed down the chain of sequences. The gate structure contains the sigmoid activation function. The Sigmoid activation function is similar to the tanh function, except that the Sigmoid compacts the value between 0 and 1 instead of between -1 and 1. This helps update or forget information because any number multiplied by 0 is 0, and the information will be removed. Similarly, any number multiplied by 1 will give you itself, and this information will be perfectly preserved. So the network knows what data needs to be forgotten and what data needs to be saved.

3.2 LASSO:

Lasso (Least Absolute Shrinkage Operator, Tibshirani (1996)) method is a compression estimation. It obtains a relatively refined model by constructing a penalty function, which compresses some coefficients and sets some coefficients to zero. Therefore, it retains the advantage of subset contraction and is biased for processing data with multicollinearity. Lasso regression model adds regularization term after loss function:

$$\frac{1}{2\mathrm{m}}[\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2+\lambda\sum_{j=1}^{n}\theta_j^2] \tag{2}$$

Where $\lambda\sum_{j=1}^{n}\theta_j^2$ is the regularization.

3.3 RF:

Random Forest (RF) is a kind of Bagging algorithm, so it is necessary to introduce Bagging algorithm before introducing Random Forest. Bagging, also known as bootstrap aggregating, is an integration technique that trains classifiers by re-selecting K new data sets from the original data set by placing back sampling. It uses the set of trained classifiers to classify new samples and then counts the classification results of all classifiers by majority voting or averaging the output. The category

with the highest result is the final label. This kind of algorithm can effectively reduce bias and variance. Compared with Bagging, RF only makes its own specification and design for some of the details.

[Weak classifier] Firstly, RF uses the CART decision tree as a weak learner. In other words, it simply refers to the Bagging method that uses the CART decision tree as a weak learner as a random forest.

[Randomness] At the same time, during the generation of each tree, the selected features of each tree are only a few randomly selected features. Generally, the square root of the total number of features m is taken by default. However, a typical CART tree will select all the features for modeling. Therefore, not only is the feature random but also the feature randomness is guaranteed.

[Sample size] Compared with the general Bagging algorithm, RF will select samples with the same number of samples as the training set N.

[Characteristics] Because of the randomness, it effectively reduces the Variance of the model, so the random forest generally does not need additional pruning. That is, it can obtain better generalization ability and Low Variance resistance. Of course, the degree of fitting to the training set will be poor. That is, the model will have a High Bias, which is only relative.

3.4 GBDT:

The decision tree used by GBDT is the CART regression tree. Because what GBDT needs to fit in each iteration is gradient value and continuous value, the decision tree used by GBDT is all CART regression tree no matter it is to deal with a regression problem, binary classification, or multiple classifications. The GDBT algorithm will be introduced in detail as follows:

(1) Initialize the learner:

$$f_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c) \tag{3}$$

(2) For m = 1,2,3,...,M are:

(a) For each sample I = 1,2,3,...,N, calculate the negative gradient, that is, the residual

$$r_{im} = -[\frac{\partial L(y_i, f_i(x_i))}{\partial f(x_i)}]_{f(x)=f_{m-1}(x)} \tag{4}$$

(b) The residual obtained in the previous step is taken as the new true value of the sample. The data $(x_i, r_{im}), i = 1,2,...,\mathrm{N}$ is taken as the training data of the next tree to obtain a new regression tree $f_m(x)$, and its corresponding leaf node region is $R_{jm}, j = 1,2,...,J$. Where J is the number of leaf nodes of the regression tree.

(c) Calculate the best-fitting value for the leaf region $j = 1,2,...,J$

$$\Upsilon_{jm} = \arg\min \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \Upsilon) \tag{5}$$

(d) Update the strong learner

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} \Upsilon_{jm} I, (x \in R_{jm}) \tag{6}$$

(3) Get the final learner

$$f(x) = f_M(x) = f_0(x) + \sum_{m=1}^{M}\sum_{j=1}^{J} \Upsilon_{jm} I, (x \in R_{jm}) \quad (7)$$

4. Model evaluation index and result analysis

The research compares the model's predicted ups and downs with the actual ups and actual downs to show how good the model's predictions are. In the training set, if the model predicts that the stock market will go up or down tomorrow, and the actual situation is the same, we call it accurate; if the model does not match, we call it deviation. The following graphs show the results obtained under each algorithm:

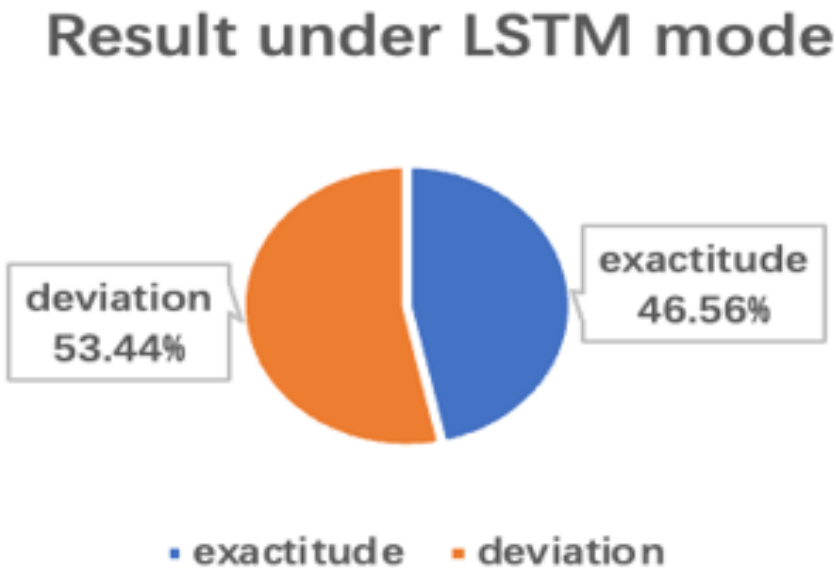

Fig 2. Comparison chart of correct and incorrect prediction result under the LSTM model.

Fig. 2 shows that under the LSTM algorithm, only 46.56% of the cases with an accurate prediction of rising or fall account for the proportion of 53.44% of the cases with a deviation of the forecast.

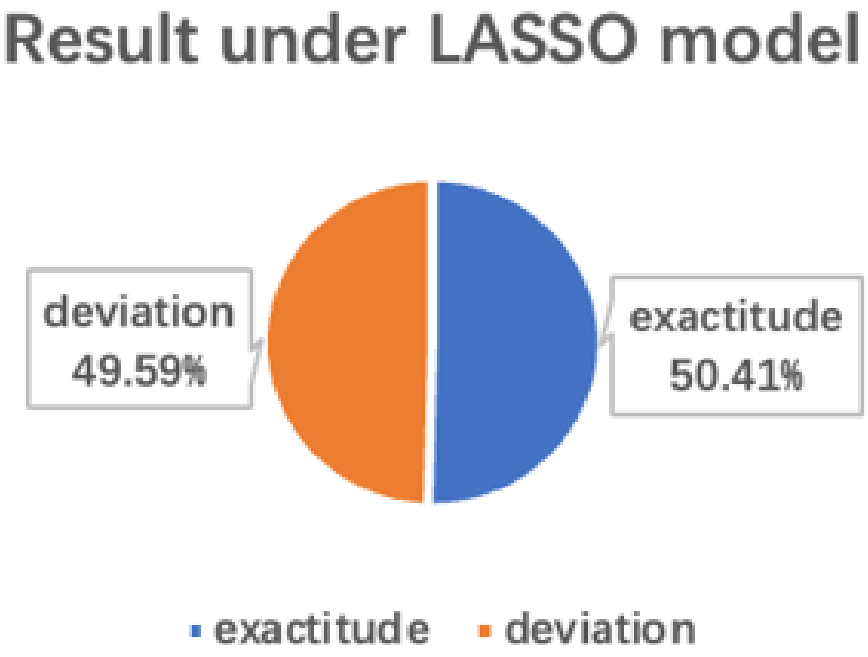

Fig 3. Comparison chart of correct and incorrect prediction result under the LASSO model.

Fig. 3 shows that under the Lasso algorithm, the percentage of the cases in which the prediction of rising or fall is accurate is 50.41%, while the percentage of the cases in which the prediction of rising or fall is wrong is 49.59%.

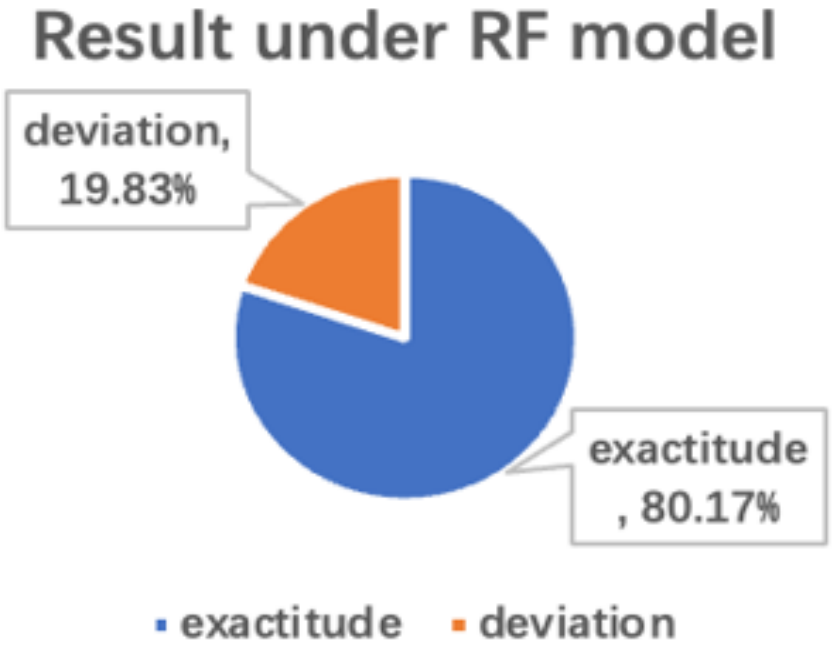

Fig 4. Comparison chart of correct and incorrect prediction result under RF model.

Figure 4 shows that under the RF algorithm, 80.17% of the cases in which the prediction of rising or fall is accurate, and 19.83% of the cases in which the prediction of rising or fall is deviant.

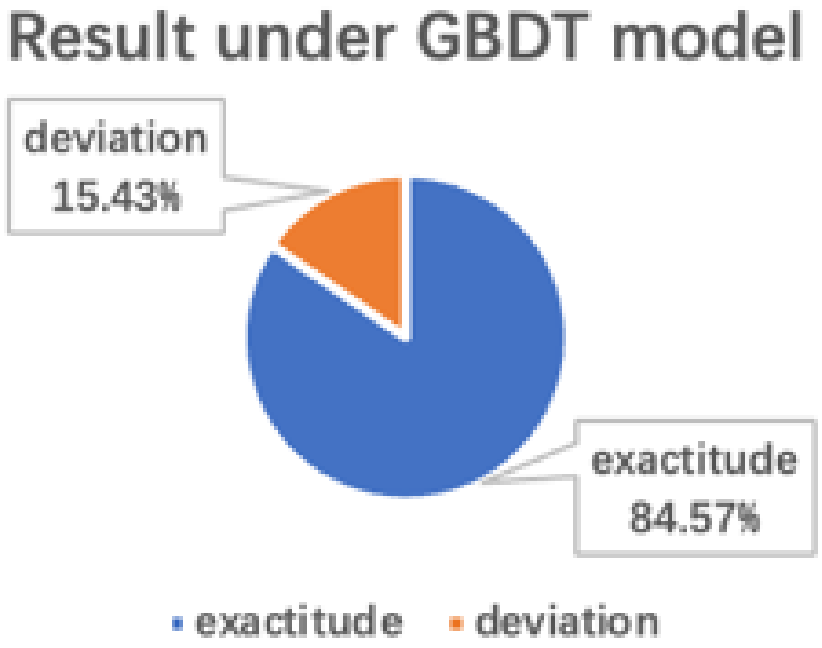

Fig 5. Comparison chart of correct and incorrect prediction result under GBDT model.

Figure 5 shows that under the GBDT algorithm, 84.57% of the cases where the prediction of rising or fall is accurate, and 15.43% of the cases where the prediction of rising or fall is deviant.

From the perspective of prediction results, the prediction effect of the LSTM model is the worst, and the accurate prediction of rising or fall is even less than 50%, and the obtained results have huge

errors. The reason for this result is that in the process of feature screening, most of the features are lagging one order. In comparison, the LSTM model needs multiple indicators lagging one order, lagging two order, or even lagging p order data, resulting in great error. The prediction under this model still has a huge error. For the Lasso model, the constraint domain is a square, and there will be a tangent point with the coordinate axis, making part of the weight zero, so it is easy to generate sparse results. Another point is that because the regularization term uses the sum of absolute values, the loss function has non-differentiable points, which leads to the failure of optimization methods such as gradient descent. Therefore, the error in the test set is very large. On the contrary, RF and GBDT models have a good prediction result, with an accuracy of 80.17% and 84.57%, respectively.

Cause effect, as well as GBDT model, is calculated for each sample will be residual, as training data under a tree, makes the study of the model is enhanced, because according to the characteristics of the selection rules, this paper selected the data basically are lagging behind the situation of the first order, so under this algorithm will have a good prediction effect.

5. Conclusion:

In this paper, investors' attention is introduced to predict the future rise and fall of the Shanghai Composite Index. Baidu index is taken as the embodiment of investors' attention, and the variable Z defined is used as the feature to select the appropriate Baidu index. The selected features are introduced into the LSTM, Lasso, RF, and GBDT models to predict the rise and fall. Due to the method of selected features and the characteristics of the model itself, through comparative analysis, we get a model with the highest proportion of prediction accuracy -- GBDT, to make a better suggestion for investors to make strategies based on this.

References

[1] S. M. Idrees, M. A. Alam and P. Agarwal, "A Prediction Approach for Stock Market Volatility Based on Time Series Data," in IEEE Access, vol. 7, pp. 17287-17298, 2019, doi: 10.1109/ACCESS.2019.2895252.

[2] Angadi M C, Kulkarni A P. Time Series Data Analysis for Stock Market Prediction using Data Mining Techniques with R[J]. International Journal of Advanced Research in Computer Science, 2015, 6(6).

[3] Wei Sun and Jingyi Sun. Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm[J]. Journal of Environmental Management, 2017, 188: 144-152.

[4] K. Chen, Y. Zhou and F. Dai, "A LSTM-based method for stock returns prediction: A case study of China stock market," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 2015, pp. 2823-2824, doi: 10.1109/BigData.2015.7364089.

[5] HYUNYOUNG CHOI and HAL VARIAN. Predicting the Present with Google Trends[J]. Economic Record, 2012, 88: 2-9.

[6] Kissan Joseph and M. Babajide Wintoki and Zelin Zhang. Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search[J]. International Journal of Forecasting, 2010, 27(4): 1116-1127.

[7] Bijl L, Kringhaug G, Molnár P, et al. Google searches and stock returns[J]. International Review of Financial Analysis, 2016, 45: 150-156.

[8] JingJing Wang. The Influence of Attention on Stock Returns: An Empirical Study of China's Securities Market [D]. Shanghai Jiao Tong University,2012