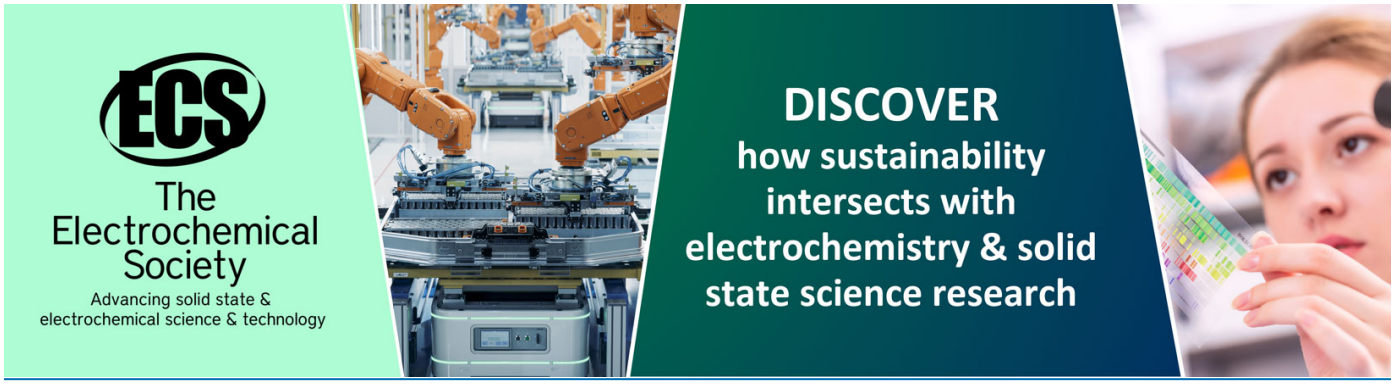
# A Quantitative Investment Model Based on Random Forest and Sentiment Analysis





# A Quantitative Investment Model Based on Random Forest and Sentiment Analysis

**Mingqin Chen, Zhenhua Zhang*, Jiawen Shen, Zhijian Deng, Jiaxing He and Shiting Huang**

Guangdong University of Foreign Studies, School of Mathematics and Statistics, 510006, Higher Education Mega Center, Panyu, Guangzhou, China.
Email: Zhenhua Zhang. zhangzhenhua@gdufs.edu.cn

**Abstract.** In recent years, under the influence of economic globalization and anti- globalization, the stock market has experienced great fluctuations in China. Quantitative investment has attracted a lot of attention because of its characteristics of maintaining stable returns. Existing research is unilaterally based on quantitative data or qualitative data for analysis to construct a quantitative investment model. This paper considers both quantitative and qualitative data to construct a more comprehensive model than that in the past. Based on the optimized database, we present a combinational model named RF-SA, which is composed of random forest and sentiment analysis model. First of all, this paper uses the SBS algorithm to select the characteristics of stock transaction historical data, optimizes the prediction database, reduces data redundancy, and improves the accuracy of the model. Secondly, we analyze the characteristics of the Chinese stock market and study the advantages and disadvantages of many data mining algorithms, and select random forest model, the most suitable model, to build the first step of stock selection model. Then, through the analysis of public opinion, the confidence index of the stockholders is calculated; on this basis, the results of the RF model and the confidence index are combined to make a second choice for the stock, and the quantitative investment portfolio is obtained, and excess returns can be obtained. The results of empirical data show that, the RF-SA model obtains a higher rate of return than the investment model of the Shanghai Stock Index.

## 1. Introduction

In recent years, due to the influence of economic globalization and anti-globalization, the stock market in China has experienced relatively large fluctuations. The performance and effectiveness of traditional investment strategies are not significant, in contrast quantitative investment strategies have attracted more and more attention with their stable investment returns and rational investment strategies.

Ross (1976) proposed an arbitrage pricing model (APT) based on investors' arbitrage behavior according to risk-return characteristics, proving that securities returns are affected by many factors. Fama and French (1993) proposed a three-factors model to classify the impact of stock returns into three factors: market asset portfolio, market value, and book-to-market ratio. On this basis, Piotroki (2000) proposed 9 indicators from profitability, liquidity and operational efficiency to establish a scoring model. In the light of Piotroki, Mohanram (2005) selected 9 indicators from profitability, growth capacity, and stability for scoring.

In recent years, many scholars have studied quantitative investment models based on data mining. Bogle et al. (2015) predicted the stock price of Jamaica through decision tree, artificial neural network and support vector machine model. Through rolling prediction method, it was found that machine

learning could achieve 90% accuracy in this market. Heatona et al. (2017) applied deep learning to the financial market for in-depth exploration, including securities pricing, portfolio management, risk control, etc. In China, Liang (2016) used the support vector machine algorithm to establish a quantitative investment stock selection model, and compared the return rate of the portfolio selected by the model with that of the Shanghai composite index. It was proved that the model could obtain a portfolio with a higher return rate than that of the Shanghai composite index. According to the CSI 300 data (Shanghai & Shenzhen), Lv et al. (2019) use the support vector machine to select short-term dominant stocks. Li et al. (2019) use a variety of machine learning algorithms to construct stock selection models, which have a higher return on investment than traditional linear regression model. Based on the traditional neural network algorithm, Xie et al. (2019) construct a long and short memory neural network model through the Bagging method of integrated learning, and obtain better results than the traditional neural network algorithm.

Although many data mining methods have been applied in the field of quantitative investment, most of these researches are based on the analysis of quantitative data, and few studies are conducted on qualitative data (e.g., shareholder comments). Zhang (2017) used the Bidirectional LSTM Model to conduct emotional analysis on the comments of investors and obtain the prediction of stock price trend. which had a high prediction accuracy in the short term, nonetheless Zhang's research did not involve quantitative analysis. Facts proved that both historical data and public sentiment exerted an influence on the trend of stocks. Nevertheless, specific studies that comprehensively consider these two factors were scarce and relevant studies were urgently needed.

Therefore, on the basis of previous studies, this paper comprehensively considers the influence of quantitative data and qualitative data on stock trend. Quantitative methods mainly focus on the analysis of turnover rate, price/ earnings ratios and stock price to predict investment income. The qualitative method mainly focuses on the sentiment analysis of the comments of investors and the calculation of their confidence index. Finally, two kinds of analysis results are synthesized to provide investment portfolio suggestions for investment institutions or individuals.

The rest of the paper is organized as follows. Section 2 introduces the concepts of SBS, RF, and lexical analysis. Section 3 introduces the steps for model construction. Section 4 focuses on experiments and result analysis. Section 5 is the conclusion.

## 2. Data Mining Algorithms and Random Forest

DT (Decision Tree), LR (Logistics Regression) and SVM (Support Vector Machine) are the most widely used in quantitative investment. As an integration algorithm, RF (Random Forest) algorithm has excellent classification performance. In order to fully compare the advantages and disadvantages of the four algorithms in quantitative investment, this paper analyzes their advantages and disadvantages from the theoretical point of view as shown in Table1.

Based on the above theoretical comparison, we believe that the RF model is the most suitable model for the following reasons:

(1) The historical data of stock is generally nonlinear and separable, so it is not suitable for LR in linear case and SVM without kernel function;

(2) RF, DT and SVM with kernel function can solve the non-linear problem. RF is an integrated algorithm of DT, hence its performance is better than DT in most cases when we deal with the non-linear data. The SVM model with kernel function is slow and difficult to determine the Super-parameters when facing the huge stock historical data. The theory of ensemble learning is used to integrate weak decision tree classifiers into robust random forest classifiers so that the generalization error of RF is smaller than that of other algorithms.

Therefore, in theory, random forest performs better than other data mining algorithms in the field of quantitative investment. This paper proposes a RF-SA quantitative investment model combining random forest model and sentiment analysis.

**Table 1.** Advantages and disadvantages of each algorithm.

| Algorithm | Advantages | Disadvantages |
|---|---|---|
| Decision Tree (DT) | (a) Decision tree model is easy to understand and implement;<br>(b) Ability to handle both data and conventional attributes;<br>(c) Insensitive to missing values;<br>(d) Uncorrelated feature data can be processed. | (a) It is difficult to predict the continuity of fields;<br>(b) When dealing with multi-classification problems, the error rate may increase faster;<br>(c) It does not perform very well when dealing with data with strong feature correlation. |
| Logistic Regression (LR) | (a) The probability of output is between 0 and 1;<br>(b) Input variables can be either continuous or categorical;<br>(c) The LR model is highly interpretable and easy to use. | (a) LR model requires more data than other models, especially sensitive to multi-collinearity.<br>(b) LR model is sensitive to abnormal values and is prone to over-fitting.<br>(c) LR model has poor performance in dealing with non-linear problems. |
| Support Vector Machines (SVM) | (a) When the amount of data is small, the SVM model will be better than other algorithms, that is, the SVM model requires less data.<br>(b) SVM model is based on the principle of structural risk minimization, which can effectively avoid over-fitting and improve generalization ability.<br>(c) SVM has better performance for high-latitude data of the model, because it is the support vector that determines the quality of its modeling, not the number of features in the training set. | (a) When the amount of data is large, the SVM model will occupy too much system memory and operation time, and the efficiency of derivation operation will be reduced.<br>(b) SVM model is sensitive to missing values, and the choice of kernel function has a great influence on the accuracy of modeling.<br>(c) Poor performance for multi-classification problems;<br>(d) It is difficult to select the optimal parameter because of the large amount of hyperparameter data after introducing the kernel function. |
| Random Forest (RF) | (a) Random forest algorithm can realize parallel operation and fast learning speed;<br>(b) These errors can be effectively balanced by random forest model with unbalanced data.<br>(c) Random forest model can well fit the non-linear decision boundary and is one of the best non-linear classification models.<br>(d) Random forest model has a strong ability to deal with missing problems, even if there are missing cases, it can still maintain a high accuracy. | (a) Random forest models are prone to over-fitting when data are noisy;<br>(b) Small data or low-dimensional data (data with fewer features) may not produce a good classification. |

## 3. RF-SA Model

### *3.1. Introduction of RF-SA Model*

The steps of RF-SA model are as follows:

**Step 1**: Data collection. We collect stock history data from RESSET database and collect stock reviews data from financial websites through web crawlers.

**Step 2**: Data pre-processing. The pre-processing of stock historical data includes data cleaning and data standardization. And the pre-processing of shareholder comment data includes word segmentation and deletion of interference items

**Step 3**: Constructing the Optimal Database. Through the SBS feature selection algorithm, we select the best feature combination and construct the optimal database with the model accuracy as the data quality evaluation standard.

**Step 4**: Constructing Random Forest model for stock selection. For the stock history data, we construct a Random Forest model to select the stock for the first time.

**Step 5**: Constructing Lexical Analysis model for stock selection. For stock reviews data, we construct a lexical analysis model to emotional analysis and calculate the Confidence Index of shareholders.

**Step 6**: Choose the best portfolio. We synthesize the results of the Random Forest model and the Lexical Analysis model to calculate the Composite Index. Based on the Composite Index, we can choose the optimal portfolio.

### *3.2. Detailed Description of RF-SA Algorithm*

**Data Preprocessing.**

Data Preprocessing of stocks historical data mainly includes outlier processing and data standardization.

(1) Detection and correction of outliers: outliers are detected by box diagram method, which are defined as values greater than $Q_U + 1.5IQR$ or less than $Q_L - 1.5IQR$. $Q_U$ is the upper quartile, which means that 1/4 of the total observed data is larger than it. $Q_L$ is the lower quartile, which means that 1/4 of the total data is smaller than it. $IQR$ is the quartile spacing, which is the difference between $Q_U$ and $Q_L$ and contains half of the observed values. The correction method adopted in this paper is to replace the values greater than $Q_U + 1.5IQR$ with $Q_U + 1.5IQR$, and the values smaller than $Q_L - 1.5IQR$ with $Q_L - 1.5IQR$.

(2) Standardization processing: This paper uses Z-score to standardize data processing. The formula is as follows:

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

Where $x$ is a specific variable, $\mu$ is the population average, $\sigma$ is the population standard-deviation.

**Sequence Backward Feature Selection Algorithms.**

Stock historical data has the characteristics of high dimensionality and large amount of data. At the same time, different stock prices may be affected by different factors. Therefore, in order to improve the speed and accuracy of the algorithm and reduce data redundancy, SBS algorithm is used to reduce the dimension of the data. The function of sequence backward feature selection algorithm (SBS) is to extract irrelevant features or noises, and automatically select the most relevant subset of features, so as to improve the computational efficiency or reduce the generalization error of the model. Under the premise of ensuring the consistency, correctness, completeness and minimization of the data, the dimensions of the data are reduced. The model is constructed as follows:

(1) Let $k = D$ initialize the algorithm, where $k$ is the number of features selected and $D$ is the dimension of feature space $X_d$;

(2) Define $x^-$ to satisfy the maximization characteristic of $x^- = \arg\max J(X_k - x)$:

(3) The feature $x^-$ is deleted from the feature set: $X_{k-1} = X_k - x^-, k = k-1$;

(4) When $k$ equals the number of targets, the algorithm terminates or jumps to step (2).

**Random Forest Quantitative Investment Model.**

In order to evaluate the quality of stocks, RF algorithm is used to construct the model with the optimal database as input, and the trend of stocks returns as output. Predicting the future return trend of the stocks by the model, divide the recommendation degree of the stocks and make the first choice. The model is constructed as follows:

(1) Using bootstrap sampling method to select "*n*" samples randomly for training set (randomly and repeatedly select "*n*" samples from the training set);

(2) A decision tree is constructed by using the selected samples in step. The rules of node partition are as follows: select "*d*" features randomly without repetition, and divide nodes according to the requirements of objective function, such as using selected features and maximizing information gain;

(3) Repeat the above process "*m*" times;

(4) Aggregate the class labels of each decision tree to vote by majority.

In this paper, the accuracy of the computational model is taken as the confidence degree of the model. The higher the accuracy of the model, the higher the credibility of the model. Because the RF models constructed by each stock are different, when the results of two models are the same, the stock with higher confidence degree will be preferred because of its higher stability and the lower risk of prediction errors.

**Constructing Sentiment Model to Select Stock.**

**Sentiment Analysis Model.**

Sentiment analysis is a process of analysis, processing, induction and inference of subjective texts with emotional colour, aiming at mining the viewpoints contained in the context. This study uses lexical analysis model to analyze emotions. The method of model is as follows:

(1) Input text and participle;

(2) Dictionary matching: using a dictionary composed of pre-marked words and lexical analyzer to convert input text into word sequence;

(3) Text matching: matching every new word with the words in the dictionary;

(4) Judging whether it matches: yes, the total score of the input text is added; if not, the score is reduced.

**Data Preprocessing.**

Word segmentation and removal of interference terms: segmenting the stock comments based the sentiment dictionary then removing interference terms to clear the text data irrelevant to the meaning of the sentence, which effectively improves the performance and accuracy of the model. The interference terms mainly include punctuation and stop words.

**Sentiment Analysis Equation.**

This paper divides stock comments into positive reviews (bullish), neutral reviews (not bullish and not bullish) and negative reviews (bearish), which have positive, non-impact and negative effects on stock prices respectively. Based on the classification results, this paper constructs a lexical analysis model to classify stockholders' comments and calculate the confidence index of each stock to measure the impact of stockholders' sentiment on stock prices. The formula is as follows:

$$The\_confidence\_index = 1\times\frac{Count(Positive\_reviews)}{Count(Valid\_reviews)} + 0\times\frac{Count(Neutral\_reviews)}{Count(Valid\_reviews)} - 1\times\frac{Count(Negative\_reviews)}{Count(Valid\_reviews)} \quad (2)$$

**Comprehensive Evaluation of the Model.**

When we calculate the model synthesis index, the effective output confidence degree obtained from the random forest model and the confidence index obtained from the lexical analysis model are considered in the following two methods:

**Method I: Simple average model.**

Ranking the confidence degree of random forest model and the confidence index of lexical analysis model respectively, then normalizing the ranking, and using the results as confidence degree score and confidence index score. The normalized formula of the confidence degree score is as follows:

$$The\_confidence\_\deg ree = \frac{The\_highest\_rank - The\_stock\_rank}{The\_highest\_rank - The\_lowest\_rank} \quad (3)$$

The formula calculates a score between 0 and 1, with the highest ranked stocks scoring 1 and the lowest scoring stocks scoring 0. The formula for calculating the synthesis index is as follows:

$$The\_synthesis\_index = The\_confidence\_index + The\_confidence\_\deg ree \quad (4)$$

The shortcoming of this method in calculating the synthesis index lies in that no matter how big the difference between the confidence degree and the confidence index, their weights are equal, ignoring the magnitude of the numerical difference between the two indexes (confidence degree and confidence index).

**Method II: Weighted average model.**

This study considers that there is a weight $\theta(\theta>0)$ between confidence degree and confidence index to make the weighted composite index conform to the following formula:

$$The\_weighted\_composite\_index = (1-\theta)\times The\_confidence\_index + \theta \times The\_confidence\_\deg ree \quad (5)$$

The disadvantage of this method is that the weight $\theta$ depends on many factors, such as investors' personal views: if investors value historical data more highly, confidence degree is more important for them and $\theta$ should take a higher value.

## 4. Example and result

All historical data of this study are from sharp database and people review data are come side wealth network shares' shareholders comments. This study selected 10 representative stock, including Shanghai Pudong Development Bank (600000), China Minsheng Banking Corp., Ltd.(600016) ,Baosteel Corporation (600019), Sinopec Group (600028), Southern Airlines (600029), Citic Securities (600030), China Merchants Bank (600036), Poly Real Estate (600048), China Unicom (600050), Shanghai Automotive Industry Corporation (600104). The empirical study was conducted on the quantitative investment and stock selection of their historical data during January 1, 2016 and July 31, 2018.

### *4.1. Determination of Input and Output Variables*

**Random Forest Model.**

**Table 2.** Output variables of RF (Random Forest).

| **Output variables** | | |
|---|---|---|
| Sequence number | Variables name | Types of variables |
| 1 | highly recommend | “2”,Classification |
| 2 | recommend | “1”,Classification |
| 3 | not recommend | “0”,Classification |

Input Variables: There 22 input variables, including Previous Close Price, Open Price, High Price, Low Price, Close Price, Adjusted Price1,Adjusted Price2,Trading Volume, Trading Sum, Daily Amplitude, Full Shares Turnover Ratio, Tradable Shares Daily Turnover Ratio, Daily Return, Daily Capital Appreciation, Equal Weighted Daily Return, Tradable Market Value Weighted Daily Return, Market Capitalization Weighted Daily Return, Equal Weighted Daily Capital Appreciation, Tradable Market Value Weighted Daily Capital Appreciation, Market Capitalization Weighted Daily Capital

Appreciation, Daily Risk Free Return, Price Earning Ratio (See Appendix I for details of the input variables). The output variables of Random Forest are shown in Table 2, and the partition standard of output variables is: "2": The yield after 30 valid trades is greater than 0 and greater than the return rate of Shanghai composite index in the same period. "1": The yield after 30 valid trades is greater than 0, but less than the return rate of Shanghai composite index in the same period. "0": The yield after 30 valid trades is less than 0.

The formula for calculating the yield after 30 valid transactions is:

$$Rate\_of\_return = \frac{Open\_price\_after\_30\_valid\_trading\_days - Current\_close\_price}{Current\_close\_price} \quad (6)$$

**Table 3.** Input variables and output variables of affective analysis model.

| Input variables | | | Output variables | | |
|---|---|---|---|---|---|
| No. | Name | Types | No. | Name | Types |
| 1 | People comment | Text | 1 | Confidence index | Numeric |

Notes: Confidence index is the number between -1 and 1, with 0 as the cut-off point, greater than 0 means bullish, less than 0 means bearish. The greater the absolute value of a confidence index, the more bullish or bearish it is.

*4.2. The Construction of Random Forest Model*

**Data Pre-processing.**

(1) Outlier Processing: In this study, outliers were corrected.

(2) Data Standardization: Standardize its variables.

**The Construction of Optimal Data Set.**

The data after data pre-processing is taken as the input of the sequential backward feature selection algorithm, and the model accuracy under different number of features is output. Thus, the minimum number of features with the highest accuracy and the corresponding optimal features are obtained.

In Figure 1,*x* axis is the number of features and *y* axis is the accuracy. Figure 1 shows the variation of accuracy as the number of features changes. In this paper, the number of features with the highest accuracy is selected for analysis. (The variation chart of SBS accuracy of other stocks is shown in Appendix II).

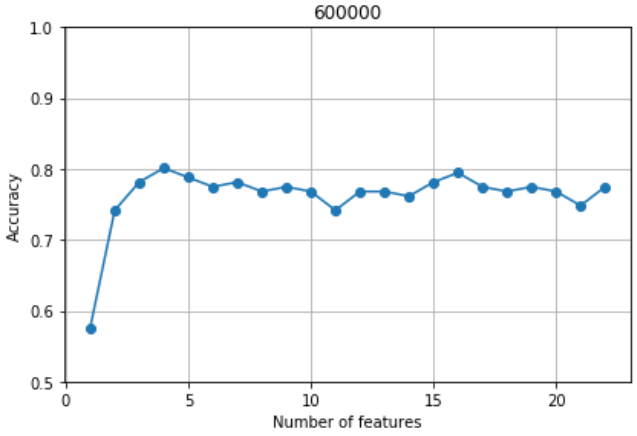



**Figure 1.** Accuracy variation diagram of SBS of Shanghai Pudong Development Bank (600000).

Do the same for the other stocks, and the results are shown in Table 4, including Previous Close Price, Open Price, High Price, Low Price, Close Price, Adjusted Price1, Adjusted Price2, Trading Volume, Trading Sum, Daily Amplitude, Full Shares Turnover Ratio, Tradable Shares Daily Turnover Ratio, Daily Return, Daily Capital Appreciation, Equal Weighted Daily Return, Tradable Market

Value Weighted Daily Return, Market Capitalization Weighted Daily Return, Equal Weighted Daily Capital Appreciation, Tradable Market Value Weighted Daily Capital Appreciation, Market Capitalization Weighted Daily Capital Appreciation, Daily Risk Free Return, Price Earning Ratio.

**Table 4.** The optimal characteristics of each stock.

| Stock code | Optimal characteristics |
|---|---|
| 600000 | 'Open Price', 'Adjusted Price2', 'Price Earning Ratio', 'Close Price' |
| 600016 | 'Low Price', 'Adjusted Price2', 'Price Earning Ratio' |
| 600019 | 'Close Price', 'Price Earning Ratio', ’Previous Close Price’ |
| 600028 | 'Adjusted Price2', 'Price Earning Ratio', 'Previous Close Price' |
| 600029 | 'High Price', 'Low Price', 'Close Price', 'Adjusted Price2', 'Trading Volume', 'Trading Sum', 'Daily Amplitude', 'Tradable Shares Daily Turnover Ratio', 'Tradable Market Value Weighted Daily Return', 'Price Earning Ratio', 'Previous Close Price' |
| 600030 | 'Adjusted Price1', 'Price Earning Ratio', 'Previous Close Price' |
| 600036 | 'Price Earning Ratio', 'Previous Close Price' |
| 600048 | 'Open Price', 'Tradable Market Value Weighted Daily Return', 'Price Earning Ratio', 'Previous Close Price' |
| 600050 | 'High Price', 'Price Earning Ratio', 'Previous Close Price' |
| 600104 | 'Adjusted Price2', 'Price Earning Ratio', 'Previous Close Price' |

**The Construction of Random Forest Model.**

Taking the optimal database as input and the recommendation degree as output, the model is trained to obtain the random forest model. Then, the random forest model of each stock is evaluated, with 20% data as the test set and 80% data as the training set to obtain the accuracy of the model.

**Model prediction.**

The established stochastic forest model is used to select stocks for the first time. Taking the data of August 1, 2018 as input, the growth of stocks after 30 trading days is predicted. The recommendation rank obtained by combining its prediction results and accuracy is shown in Table 5. According to the results of recommendation degree output, five stocks can be preliminarily selected through random forest, and their comprehensive recommendation degree ranking can be obtained through the accuracy.

**Table 5.** Each stock comprehensive recommendation degree ranking.

| Stock code | Prediction results | Accuracy | Ranking |
|---|---|---|---|
| 600050 | 2 | 0.8584 | 1 |
| 600028 | 2 | 0.808 | 2 |
| 600000 | 2 | 0.7933 | 3 |
| 600019 | 2 | 0.7924 | 4 |
| 600029 | 2 | 0.7419 | 5 |
| 600104 | 0 | 0.808 | 6 |
| 600048 | 0 | 0.8048 | 7 |
| 600036 | 0 | 0.8 | 8 |
| 600030 | 0 | 0.7933 | 9 |
| 600016 | 0 | 0.76 | 10 |

**Table 6.** The results of DT (Decision Tree Model).

| Stock Code | Prediction results | Accuracy | Ranking |
|---|---|---|---|
| 600050 | 2 | 0.8208 | 1 |
| 600029 | 2 | 0.7419 | 2 |
| 600016 | 2 | 0.616 | 3 |
| 600000 | 1 | 0.7025 | 4 |
| 600019 | 0 | 0.8396 | 5 |
| 600104 | 0 | 0.8 | 6 |
| 600028 | 0 | 0.792 | 7 |
| 600030 | 0 | 0.7655 | 8 |
| 600036 | 0 | 0.76 | 9 |
| 600048 | 0 | 0.7561 | 10 |

**Table 7.** The results of LR(Logistic Regression).

| Stock Code | Prediction results | Accuracy | Ranking |
|---|---|---|---|
| 600019 | 2 | 0.6792 | 1 |
| 600028 | 2 | 0.648 | 2 |
| 600029 | 2 | 0.6371 | 3 |
| 600016 | 2 | 0.632 | 4 |
| 600000 | 2 | 0.5455 | 5 |
| 600036 | 0 | 0.68 | 6 |
| 600050 | 0 | 0.6604 | 7 |
| 600048 | 0 | 0.6423 | 8 |
| 600104 | 0 | 0.6 | 9 |
| 600030 | 0 | 0.5868 | 10 |

**Table 8.** The results of SVM.

| Stock Code | Prediction results | Accuracy | Ranking |
|---|---|---|---|
| 600000 | 2 | 0.7355 | 1 |
| 600029 | 2 | 0.7097 | 2 |
| 600028 | 2 | 0.704 | 3 |
| 600030 | 2 | 0.6942 | 4 |
| 600036 | 2 | 0.664 | 5 |
| 600016 | 2 | 0.648 | 6 |
| 600104 | 2 | 0.568 | 7 |
| 600048 | 0 | 0.7154 | 8 |
| 600050 | 0 | 0.7075 | 9 |
| 600019 | 0 | 0.6792 | 10 |

**Comparison between LR, SVM, DT and RF Model.**

**Table** 6, Table 7, and Table 8 show the predicting results of DT, LR and SVM.

In the following content we will choose the most appropriate algorithm model from the algorithms that we have mentioned before by comparing the accuracy of their results.

**Table 9.** The average confidence level of RF, DT, SVM, and LR.

| Model | Average accuracy |
|---|---|
| Random Forest Model | 0.79601 |
| Decision Tree Model | 0.75944 |
| SVM | 0.68255 |
| Logistic Regression | 0.63113 |

From Table 9 we can see that the average confidence level of Random Forest Model is the highest, which means that Random Forest Model is more stable and less error making than the other models. That makes Random Forest Model better and more accurate than the other models. In this case, we can make a conclusion that Random Forest Model is the most suitable for predicting how stocks' trend will go.

*4.3. Emotional Analysis Model*

**Data pre-processing.**

Firstly, collect the text using a spider program and use Jieba, an open source word segmentation tool, to achieve word segmentation. Jieba has an active community, abundant functions and it can be simply utilized. It can complete the segmentation work conveniently and quickly. In this paper, we adopt the precise word segmentation mode of Jieba segmentation tool, which is suitable for text analysis and can help cut sentences more precisely.

Next, to improve the data quality, we delete the interference words using the Stopword list of Harbin Institute of Technology together with some meaningless words which are proprietary in stock reviews in the process of identifying the interference terms.

**Construction of Lexical Analysis Model.**

We construct a lexical analysis model with the pre-processed data used as input variables and confidence index as output variables. Then, we construct a suitable dictionary with some special emotional words such as "Bei Tao"(which means "trapped in the stock market" in Chinese), "Jia Cang"(which means "buy in" in Chinese) and "Bi Hong"(which means "on fire" in Chinese) on the basis of Tsinghua University's Chinese Dictionary of praise and derogation The results of the model are shown in Table 10.

**Table 10.** Confidence index and ranking of samples.

| Stock Code | Confidence index | Ranking |
|---|---|---|
| 600028 | -0.0078 | 1 |
| 600050 | -0.0596 | 2 |
| 600029 | -0.2047 | 3 |
| 600019 | -0.2080 | 4 |
| 600000 | -0.4168 | 5 |

### *4.4. Comprehensive Model Results Analysis*

**Method I: Simple average model.**

Based on the Random Forest model (Model 1) and Lexical Analysis model (Model 2), the results of the synthesis index constructed by Method I are shown in Table11.

**Table 11.** Comprehensive index and ranking of each stock by the Method I.

| Stock code | Confidence degree ranking | Confidence index ranking | Synthesis index | Synthesis index ranking |
|---|---|---|---|---|
| 600028 | 2 | 1 | 1.75 | 1 |
| 600050 | 1 | 2 | 1.75 | 1 |
| 600000 | 3 | 5 | 0.5 | 3 |
| 600019 | 4 | 4 | 0.5 | 3 |
| 600029 | 5 | 3 | 0.5 | 3 |

**Method II: Weighted average model.**

Assuming a uniform distribution of 0 to 1: $\theta \square U(0,1)$ , that is, when the intervals of confidence degree and confidence index are 0 to 1, the results of the weighted composite index are calculated by method II as shown in Figure 2.

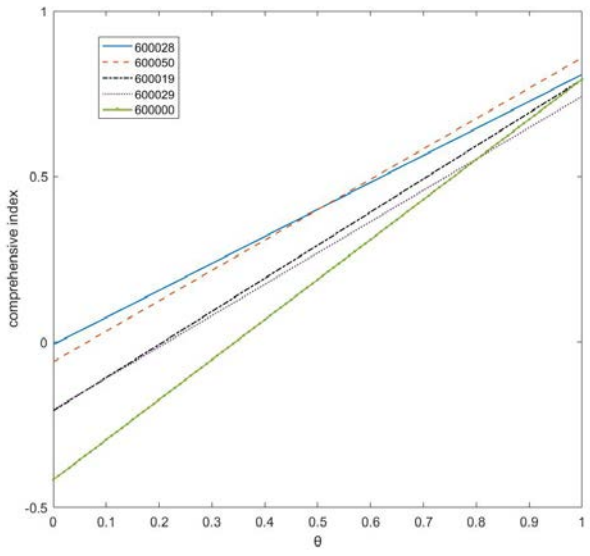



**Figure 2.** The results of weighted composite index for $\theta$ in [0,1].

**Table 12.** The ranking of weighted composite index calculated by Method II.

| Stock code | Confidence degree | Confidence index | Weighted composite index | Weighted index ranking |
|---|---|---|---|---|
| 600028 | 0.808 | -0.0078 | 0.4000 | 1 |
| 600050 | 0.8584 | -0.0596 | 0.3993 | 2 |
| 600019 | 0.7924 | -0.2080 | 0.2921 | 3 |
| 600029 | 0.7419 | -0.2047 | 0.2685 | 4 |
| 600000 | 0.7933 | -0.4168 | 0.1882 | 5 |

The following formula to calculate the expectation of composite index:

$$EX = \int_0^1 (a\theta + b(1-\theta))d\theta = \frac{a+b}{2} \tag{7}$$

Where $a$ is the confidence degree, $b$ is the confidence index. The results are shown in Table 12.

In Table 12, the values the weighted composite index of Sinopec (600028) and China Unicom (600050) are significantly higher than the other stocks. And based on the results of random forest model and public sentiment analysis, the optimal investment portfolio is Sinopec (600028) and China Unicom (600050). The difference between the two methods is that Method I will have apposition situation because it scores using the same index in the case of fewer stocks to be selected.

Evaluate the effect of investment through historical data. Assuming that the holding period is from August 1 to August 31, 2018, the holding period yield is calculated as shown in Table 13.

**Table 13.** Comparison of the return between the RF-SA model and the Shanghai Stock Index.

| Investment choices | Holding period yield |
|---|---|
| RF-SA Investment Portfolio(Two Stocks)(%) | 5.4191 |
| The First Stocks Selection(RF algorithm only, Five Stocks)(%) | 0.0081 |
| All Alternative Stock(Ten Stocks)(%) | -0.0085 |
| Shanghai Stock Exchange Composite Index(%) | -3.51505 |

We can see from Table 13 that during the period, the investment portfolio has achieved higher returns compared with Shanghai Stock Exchange Composite Index. The Shanghai Stock Exchange Index reflects the overall situation of all stocks on the Shanghai Stock Exchange. The negative return indicates that the stock price has a downward trend. In this case, the model can still choose stocks with higher returns to build a portfolio to ensure investment returns.

## 5. Conclusions

This paper presents a new quantitative investment model which combines quantitative analysis with qualitative analysis. Firstly, based on historical data, the random forest model in data mining is used for quantitative analysis. At the same time, it introduces the method of lexical analysis to conduct qualitative analysis, excavates useful information from the comments of shareholders, and constructs the confidence index of shareholders. Finally, a weighted composite evaluation model is established for the results of quantitative and qualitative analysis, which makes up for the deficiency of quantitative analysis or qualitative analysis. The empirical results show that, by combining the random forest stock selection model and the public opinion analysis model, the comprehensive stock selection model can obtain a portfolio that is higher than the return rate of Shanghai composite index (i.e., excess return), which is in line with the expectation. Due to the limitation of time, energy and funds, the existing research is still not perfect and needs further exploration. Firstly, only a limited number of 10 stocks were simulated for stock selection in this paper, failing to test stock selection under more complex conditions. Subsequent studies will consider relaxing stock selection pools to test and improve the performance of models in complex market environments. Secondly, this study did not include some quarterly financial indicators of the company as features in the model, and the index selection can be relaxed later, so as to find more representative features for the construction of the optimal database. Third, the random forest model and lexical analysis model used in this paper are both static models, and the subsequent research will consider the use of dynamic model, that is, the sample will be added to the model every time the prediction is made to realize the dynamic update of the model, so as to increase the training sample and complexity of the model and obtain a more accurate prediction model. Finally, our research on the analysis of public opinion is not deep enough. We for all categories of people over a period of time the effective comments are analyzed, in fact, the public opinion analysis can be divided into big V (post reading more and more people convincing) comments, V (post reading quantity is more, some people convinced) comments, and Volkswagen (post read less) comments, different people may have different effects on the stock market by public

opinion. We believe that with the rapid development of the Internet today, this research is very necessary. In future studies, we will carefully study the influence of different opinions on the financial market to make our model more accurate.

## 6. Acknowledgements

This paper is funded by the National Statistical Research Key Project (No.2016LZ18), Guangdong Basic and Applied Basic Research Foundations (No.2018A030313470, 2016A030313688), Student Innovation Projects (No. S201911846035, S201911846029, 201511846058) of Guangdong Province, Teacher-Student Joint Research Project (No.18SS08) & Higher Education Project (No.2019GJ13Y) of Guangdong University of Foreign Studies.

## 7. References

[1] Stephen A Ross 1976 The arbitrage theory of capital asset pricing *Journal of Economic Theory* vol 13(3) p 341-360
[2] Eugene F Fama and Kenneth R French 1993 Common risk factors in the returns on stocks and bonds *Journal of Financial Economics* vol 33(1) p 3-56
[3] Piotroski J D 2000 Value investing: The Use of Historical Financial Statement Information to Separate Winners Form Losers *Journal of Accounting Research* vol 38(2) p 1-41
[4] Partsa S Mohanram 2005 Separating Winners from Losers among Low Book-to Market Stocks using Financial Statement Analysis *Review of Accounting Studies* vol 10(2) p 133-170
[5] Bogle S A, Member and IAENG W D Potter 2015 A machine learning predictive model for the Jamaica Frontier Market *Proceedings of the World Congress on Engineering WCE 2015 London* vol 1 p 1-6
[6] Heatona J B, Polsonb N G and Wittec J H 2019 Deep learning for finance: deep portfolios *Applied Stochastic Models in Business and Industry* vol 33(1) p 3-12
[7] Liang C 2016 *Research on Adaptability of Quantitative Investment Model - Analysis of quantitative stock selection model based on support vector machine* (Beijing: Central University of Finance and Economics)
[8] Lv K C, Yan H F and Chen C 2019 Quantitative Investment Strategy Based on CSI 300 *Journal of Guangxi Normal University (Natural Science Edition)* vol 37(01) p 1-12
[9] Li B, Shao X Y and Li Y Y 2019 Research on Machine Learning Driven Quantamental Investing *China Industrial Economics* vol 8 p 61-79
[10] Xie Q, Cheng G G and Xu X 2019 Research Based on Stock Predicting Model of Neural Networks Ensemble Learning *Computer Engineering and Applications* vol 55(08) p 238-243
[11] Zhang K, Ren W P, Zhang Y S and You J Q 2017 Research on stock prediction method based on information of investors' comments *Journal of Beijing Information Science & Technology University (Natural Science Edition)* vol 32(05) p 67-71