Eng Int Syst (2022) 6: 431–440
© 2022 CRL Publishing Ltd

**Engineering**
*Intelligent Systems*

# Impact of Online Stock Reviews on Stock Market Trends Using Text Sentiment Analysis Methods

**Haixiang Li[1,2,*] and Weixian Xue[1]**

[1] *School of Economics and Management, Xi'an University of Technology, Xi'an 710048, Shaanxi, China*
[2] *School of Humanities and Management, Xi'an Traffic Engineering Institute, Xi'an 710300, Shaanxi, China*

As an important part of the financial market, the stock exchange market plays a very important role in the development of the economy. To better predict stock trends, this article proposes a text sentiment analysis method based on the influence of online stock reviews on stock market trends. This article takes online comments (posts) in Internet stock forums as the research object, including the amount of comment information and the emotional tendency expressed by specific content. In this paper, Chinese documents are processed for word segmentation. According to certain feature selection and feature extraction methods, feature words are selected to be counted as a feature word matrix, and the corresponding feature word frequency is calculated. The word segmentation tool used in this article is the Rwordseg package provided by $R$, through which the article can be easily segmented to form a word frequency matrix. First, words are selected from the suggested word bank to match the second half of the stock review. If the matching is unsuccessful, the following words continue to be matched until the matching is successful and the matched words are used as the suggested words. This article divides the views of stock investors into rise, neutral and fall, and uses the trading day as the time unit to conduct sentiment analysis on the text data. While analyzing the weight of the stock review authors, this article also analyzes the attention degree of the stock reviews. The predicted value of the multi-data source stock prediction model of the simple sentiment index is closer to the actual closing price, the relative error between the predicted value and the true value is no more than 1.9%, and the prediction trend accuracy rate is 78.9%, which does not include the simple sentiment index. The relative error between the predicted value and the true value of the multi-data source stock prediction model is 6.3%. The results show that online stock reviews affect the trend of stocks, positive sentiment causes stocks to rise slightly, and negative sentiment causes stocks to fall.

Keywords: Online Stock Reviews, Stock Market Trends, Text Sentiment Analysis, Stock Forecasts

## 1. INTRODUCTION

In recent years, in the context of the continuous development of machine learning, data mining and other related fields, computer algorithms are used to conduct in-depth mining and analyse the capital market to explore the law of stock rise and fall. As a trading market with the deep participation of investors, the stock market is not only affected by macro and micro economic factors, but also by changes in investor sentiment and the psychological expectations of the stock market. Compared with the real economy, investors have a certain degree of speculation in stock transactions. Therefore, changes in the stock market will also have a great impact on investors' psychology.

This paper reports on research into text big data processing system architecture, analyzes the key technologies needed to build investor sentiment indicators, and combined with the research of key technologies of stock index data calculation and stock trend prediction model, describes the design a stock prediction system supporting big data processing, which provides a solution for investors to make scientific decisions and effective investments. This is helpful to understand how investor sentiment, as the main component of behavioral

[*] Address for correspondence: Haixiang Li, School of Economics and Management, Xi'an University of Technology, Xi'an 710048, Shaanxi, China, Email: lihaixiang_1986@163.com.

finance, affects the behavior characteristics of investors, master the behavior and psychological characteristics of investors and their relationship with the operation of the stock market, and provides a reference for individual investors when they invest.

Discussions on stocks on the Internet can significantly affect their price trend. Huang [1] states that conducting sentiment analysis on social media data is key to understanding people's positions, attitudes and opinions on a certain event, and it has a wide range of applications such as election forecasting and product evaluation. Although researchers have worked hard on a single modality (image or text), they have paid less attention to the joint analysis of multimodal data in social media. Huang proposed a new image text sentiment analysis model, namely Deep Multimodal Attention Fusion (DMAF). He proposed two independent unimodal attention models, which are effective emotion classifiers for learning visual and text modality respectively. Then, he proposed a multimodal attention model based on intermediate fusion, which uses the inherent correlation between visual features and text features to perform sentiment joint classification. Finally, he adopted a new fusion scheme to combine the three attention models for emotion prediction. Although his research is more effective on weakly labeled and manually labeled data sets, it lacks the necessary experiments to prove his point of view. Singh [2] proposed a novel verb and spell check system framework which extracts user reactions, emotions and opinions from social media text (SMT). He discussed the various steps of the framework, such as lowercase conversion, tokenization, spell checking, part-of-speech tagging, stop word elimination, stemming, sentiment score calculation, and SMT classification. He first proposed the concept of threshold negative parameters. In the experiment, he evaluated the performance of the system on three data sets, namely Facebook's comments on India's Goods and Services Tax (GST) implementation, Twitter debates between former US Presidents Obama and McCain, and movie reviews. Although his system performance is better than other methods, it lacks innovation. Xu [3] found that the rapid growth of web data volume presents a severe challenge to web monitoring and that the accuracy of traditional text sentiment analysis methods may be reduced when dealing with massive amounts of data. After the training process, the eigenvalues of the CNN are unevenly distributed. To overcome this problem, he proposed a method of eigenvalue normalization. Through simulation, he optimized the size of the text features. Finally, he proposed a method to update the learning rate during CNN training to achieve better results.. The purpose of Yu [4] research is to explain how to mine open answers and emotional expressions (positive or negative) from student surveys and provide valuable information to improve student experience management (SEM). With the development of text mining technology based on artificial intelligence and machine learning, previously underutilized text data has been found to have an important application value in SEM. To illustrate how to apply text mining to SEM, he discussed an example of a campus survey conducted by Arizona State University. Instead of imposing the researcher's preconceived assumptions on the students using the forced option survey project, he chose to use open-ended questions to elicit a freely appearing theme from the students. Although his research methods are relatively comprehensive, they lack specific experimental content [4].

This research uses text sentiment as the entry point, extracts investor sentiment information from online public opinion, analyzes its impact on the stock price of the GEM, and helps the relevant authorities to correctly grasp the overall sentiment of investors and fully understand the influencing factors of the stock price fluctuation of the GEM, and then advances its adoption of policy measures to stabilize the stock market by providing a reference technical method, and also provides new ideas for related research on behavioral finance.

## 2. ONLINE STOCK REVIEWS AND TEXT SENTIMENT ANALYSIS

### 2.1 Online Stock Review

Assuming that the total number of texts is $N$, the number of texts containing the feature word $t_i$ is $n_i$, the frequency of the feature word $t_i$ in text $D$ is $f_i$, the inverse text frequency $idf_i$ of the characteristic word $t_i$ is expressed as $\log(N/n_i)$, and the calculation formula of the characteristic word $t_i$ is:

$$w_i = \frac{f_i \times \log(N/n_i + 1)}{\sqrt{\sum_{t_i \in D} [f_i \times \log(N/n_i + 1)]^2}} \tag{1}$$

where $w_i$ refers to the weight of feature words $t_i$ in text D [5].

The content of a concept word can be expressed by calculating the frequency of its appearance in the document. If the frequency of occurrence is high, this means that the content is rich; otherwise, it means that the content is lacking. The calculation formula is [6]:

$$P(w) = \frac{\sum\limits_{n \in word(w)} n}{N} \tag{2}$$

where $P(w)$ refers to the frequency of the concept word $w$, and $N$ refers to the total number of concept words.

Assuming $S(w_1, w_2)$ is the set of concept words, $w_1$ and $w_2$ are the common parent node concept words, then the formula for calculating the common parent node is [7]:

$$P_{\min}(w_1, w_2) = \min_{w \in S(w_1, w_2)}\{P(w)\} \tag{3}$$

Then the similarity calculation formula of concept words $w_1$ and $w_2$ is:

$$Sim(w_1, w_2) = 1 - P_{\min}(w_1, w_2) \tag{4}$$

If the self-contained content of the concept words $w_1$ and $w_2$ is taken into account, the calculation formula is [8]:

$$Sim(w_1, w_2) = \frac{2 \ln p(w_1, w_2)}{\ln p(w_1) + \ln p(w_2)} \tag{5}$$

TF-IDF is specifically used to evaluate the importance of a word in the entire document set. Term frequency (TF) refers to the number of times a feature item appears in the document,

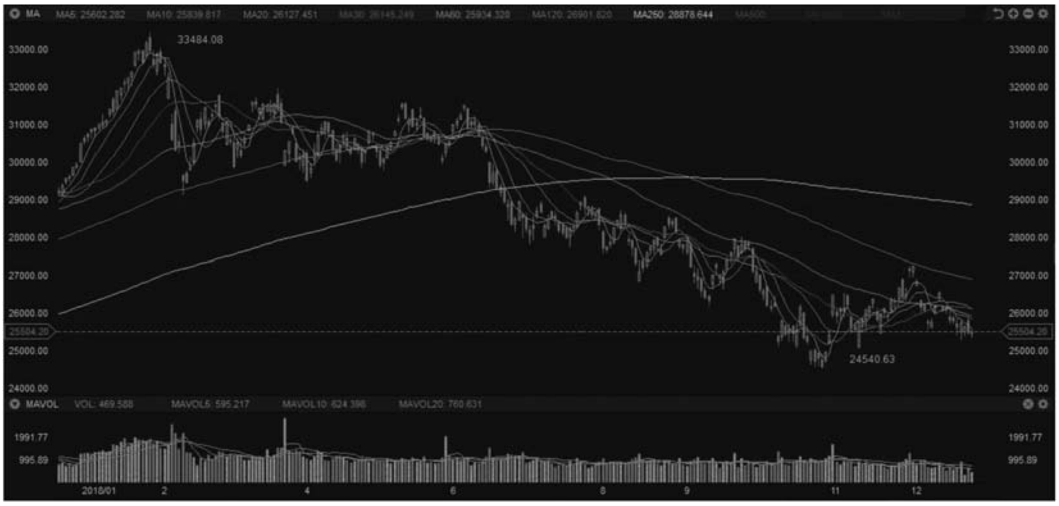

**Figure 1** Stock trend (picture source: http://alturl.com/jfb5h).

reflecting the importance of a feature item to the text. The expression is as follows [9–10]:

$$W_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log\left(\frac{N}{n_j}\right) \tag{6}$$

where $n_j$ is the number of documents with feature items, $N$ is the total number of documents, and $idf_j$ is the reciprocal of documents with feature items [11].

Assume that the time series $y_t$ satisfies the following ARFIMA(p,d,q) model [12]:

$$\phi(B)(1-B)^d(y_t - u) = \psi(B)\varepsilon_t \tag{7}$$

Among them,

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p \tag{8}$$

$$\psi(B) = 1 + \psi_1 B + \ldots + \psi_p B^p \tag{9}$$

where $B$ is the lag operator, $\varepsilon_t$ is the independent and identical distribution satisfying the mean value of zero, and $u$ is the mean value of the sequence $y_t$ [13].

For any two volatility models u, v, the relative loss function can be calculated, denoted as $d_{i,uv,m}$. The formula is as follows:

$$d_{i,uv,m} = L_{i,u,m} - L_{i,v,m}(i, j = 1, \ldots, m; t = 1, \ldots, n) \tag{10}$$

where $L_{i,u,m}$, $L_{i,v,m}$ are the loss function values calculated by model u and model v using the predicted value of the volatility of the next day under a certain loss function [14–15].

## 2.2 Stock Market Trends

The normalization of numerical data can ensure that the range of features is roughly within a certain fixed range, in order to eliminate the dimensional influence between indicators, data standardization processing is required to solve the comparability between data indicators. Secondly, after the data is processed, it can ensure that the classifier is more efficient in classifying and avoids a reduction in the learning speed of the classifier because the dimensionality of the input data is too large [16]. The calculation formula of the minimum and maximum method is:

$$origin_i^* = (y_{\max} - y_{\min}) \times \frac{origin_i - origin_{\min}}{origin_{\max} - origin_{\min}} + y_{\min} \tag{11}$$

where $origin_{\max}$ represents the maximum value of the original data set, and $origin_{\min}$ represents the minimum value of the original data set. $y_{\max}$ represents the maximum value after data normalization, and $y_{\min}$ represents the minimum value after data normalization [17].

Let $S_i$ be the number of samples in the class $C_i$. The expected information required to classify a given sample is as follows:

$$I(S_1, s_2, \ldots, s_m) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{12}$$

where $p_i$ is the probability that any sample belongs to $Ci$, and it is estimated with $s_i/s$ [18].

After using the LDA topic extraction model to extract topics from the daily stock review data, a large amount of stock review data is integrated into the form of "topics", and each topic text of each stock is analyzed by date, and the concept of the public opinion index (POI) is introduced. It is used to measure the quantified topic information. The product of the sentiment value of each topic and the weight of the feature term is used to represent the POI of the topic, and then the POI of all topics is summed to obtain the stock corresponding to that day. The POI is calculated as follows:

$$POI = \sum_{k=1}^{K}\left(e_k \cdot \sum_{n=1}^{N} Weight(W_{kn})\right) \tag{13}$$

where $k = \{1, 2, \ldots, K\}$ represents the number of topics in the daily stock review text; $e_k$ represents the sentiment value of the $k$-th topic, and the result is classified by the text classifier [19].

The stock trend is shown in Figure 1. The opinions and attitudes expressed by investors on stock market trends have a certain emotional tendency, which may have an impact on

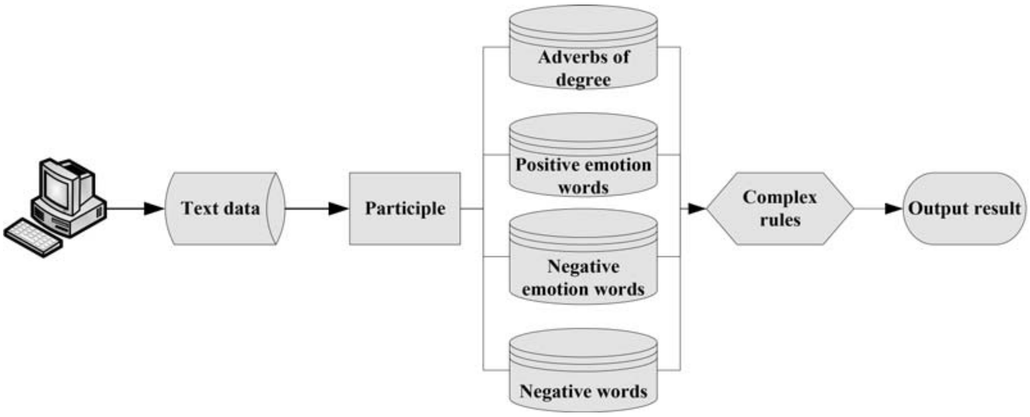

**Figure 2** Rule-based sentiment analysis model process.

the trading situation of the entire stock market. Financial news and other information may change their judgments and behavioral decisions by influencing investor sentiment, and ultimately affect the volatility of the stock market [20].

## 2.3 Text Sentiment Analysis

To compare the time series of public sentiment value and the time series of stock market-related variables, it is necessary to match the daily sentiment value and stock market-related variables. Since stock bars generate public sentiment information almost every day, the stock market trading hours are 9:30–11:30 in the morning and 13:00–15:00 in the afternoon from Monday to Friday.. In this case, it is necessary to focus on the information of the stock market and remove the public sentiment information on a stock market holiday. This is obviously a waste of public sentiment information obtained through complex calculations; or to supplement the relevant information on the stock market holiday, so that there is stock market information corresponding to the daily public sentiment value [21–22].

The process of the rule-based sentiment analysis model is shown in Figure 2. A sentiment dictionary generally comprises four parts: degree adverb dictionary, positive emotion dictionary, negative emotion dictionary and negative word dictionary. Due to the non-linear characteristics of stock prices and the characteristics of high noise in stock market data, it is difficult to make more accurate predictions using only stock data analysis. In today's Internet era, a large number of users publish their opinions and comments on the stock market in various stock forums every day, resulting in a large number of online texts with great research value. This information often contains investor feedback and relevant comments on the stock market and investment plan information. Obtaining this information is very helpful to the study of future stock trends [23].

In reality, text data is the main form of information. The demand to extract useful knowledge for scientific research and business activities from text data is gradually increasing. In stock market research, text mining methods are used to extract information about the stock market trends from text information related to stocks, such as stock news, online stock reviews, etc., to provide new factors for stock market prediction and analysis. There are qualitative differences between the positive and negative inertia measured by a person over time. The inertia of positive influence only gradually increases with time, while the negative influence gradually changes with time. This illustrates the need to be able to model gradual changes and sudden changes in order to detect meaningful quantitative and qualitative differences in temporal emotions [24].

Investor sentiment reflects the mentality changes and behavioral characteristics of market investors in the investment process. Investor sentiment research is a hot issue in behavioral finance. Although there are differences in the preferences of different investors, under mutual learning, imitation and influence, the emotional characteristics and investment behaviors of investors will also tend to be consistent, which leads to financial asset prices deviating from the true value.

Each language model represents a certain emotional tendency. In text classification, the pre-processed text to be classified is matched with the pattern in the pattern library, and the emotional value represented by each pattern is accumulated to express the emotion of the text. The process of text representation is through the preprocessing of the original text. After the steps of word segmentation, part-of-speech tagging, and removal of stop words, the relevant feature selection methods such as document frequency, information gain, mutual information, part-of-speech position, etc. are used to extract the text features and the feature weights are calculated using statistical methods. Finally, the feature weights are normalized so that the input and output meet the requirements of classification algorithms. With the arrival of the information age, a large number of network terms and other unconventional language forms continue to appear. This not only necessitates new requirements for word-level text sentiment analysis, it also allows scholars, especially researchers in computational linguistics, to use fine-grained sentiment analysis research results and methods to deal with more complex phrases containing multiple components. Sentiment analysis can be performed with sentences and various weight calculation methods can be used to analyze the sentiment tendency of the entire paragraph or the entire document [25].

**Table 1** Statistical results of rational and irrational sentiments of individual and institutional investors.

| Emotional variables | Mean | Standard deviation | Skewness | Kurtosis |
|---|---|---|---|---|
| Institutional rationality | – 1.0573 | 4.6862 | – 1.3170 | 9.3225 |
| Institutional irrational emotions | 3.15e-09 | 5.4784 | 0.7704 | 6.6062 |
| Personal rational emotion | 3.1789 | 3.2236 | 1.4100 | 9.5673 |
| Personal irrational emotions | 3.00e-09 | 2.5506 | – 0.8380 | 10.4157 |

# 3. STOCK TREND PREDICTION EXPERIMENT

## 3.1 Data Sources

This research uses online comments (posts) in Internet stock forums as the research object, including the amount of comment information and the emotional tendency expressed by the specific content. All types of websites summarize the listed companies according to the stock name or stock code to facilitate data queries. Each stock has its own discussion section, so there is no difference between various websites in terms of a data search.

## 3.2 Data Preprocessing

In this study, Chinese documents are processed for word segmentation. According to certain feature selection and feature extraction methods, feature words are selected to be counted as a feature word matrix, and the corresponding feature word frequency is calculated. The word segmentation tool used in this article is the Rwordseg package provided by R, through which the article can easily be segmented to form a word frequency matrix. In this research, multiple training sets are constructed using the random under-sampling method. Each training set consists of the remaining few-class samples and the multi-class samples obtained by random under-sampling. Different base classifiers are trained on different training sets and these different base classifiers are used to classify the same test set. Finally the classification results of each base classifier are combined.

## 3.3 Feature Word Extraction

A certain word in the text can characterize the classification characteristics of the document to a certain extent, but different words have different characterization capabilities for the text category, and this characterization ability is generally represented by statistical probability. A value can be set as a threshold for the probability of the classification and the characterization ability of the word. The characterization ability of all words in the text segmentation result after removing the stop words is calculated and the characterization ability is found to be small. For the feature extraction result, the characterization ability is found to be sufficiently large. After the feature word matching is completed, the suggested word can be matched. First, words are selected from the suggested word bank to match the second half of the stock review. If the matching is unsuccessful, the following words continue to be matched until the matching is successful and the matched words are used as the suggested words.

## 3.4 Construction of Sentiment Indicators

In this research a direct indicator, the Consumer Information Index (CCI) is selected and five indirect indicators, turnover rate (TURN), number of new accounts (ACCN), price-earnings ratio (PE), volume (VOL), and public opinion index (TEXT) are also selected. A total of six variables are used to construct a comprehensive index of investor sentiment. In addition to its own influence on investor sentiment, it also has its lag, so a total of twelve variables are used for principal component analysis with the lag of these six variables in one period.

## 3.5 Sentiment Analysis

In this study, the views of stock investors are divided into rise, neutral and fall. The sentiment analysis of the text data is carried out with the trading day as the time unit, and the statistics are the stock review text data on the stock trading day. In the article, when we set the sentiment tendency judgment index, the main analysis is the number of texts of bullish and bearish views. According to the results of the text classification of the specific stocks in each statistical time period, the sentiment tendency index and the sentiment polarization index of each time period are obtained. When analyzing the weight of the stock review authors, this article also analyzes the attention degree of the stock reviews. The amount of reading and commenting on the stock review reflects to a certain extent the attention of investors to this stock review, and to a certain extent can reflect the influence of the opinions of the stock review. However, when the influence weight index is set, the weight distribution of stock review reading volume and comment volume is not balanced.

# 4. RESULTS AND DISCUSSION

The statistical results of the rational and irrational emotions of individual and institutional investors are shown in Table 1. We find that institutional investors' rational emotions are less volatile than their irrational emotions. This is because rational emotions are supported by economic fundamentals, while non-rational emotions are influenced by the psychological changes of investors, which are different for each participant, hence the fluctuation for irrational emotions is relatively large. But the rational sentiments of individual investors fluctuate

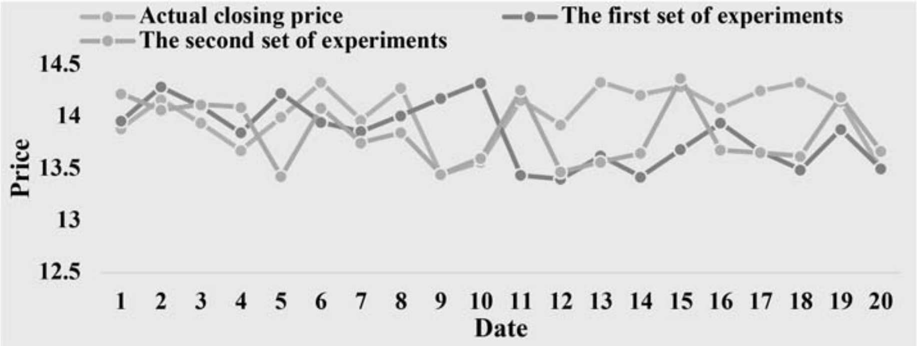

**Figure 3** Comparison of prediction results and actual values of stock prediction models.

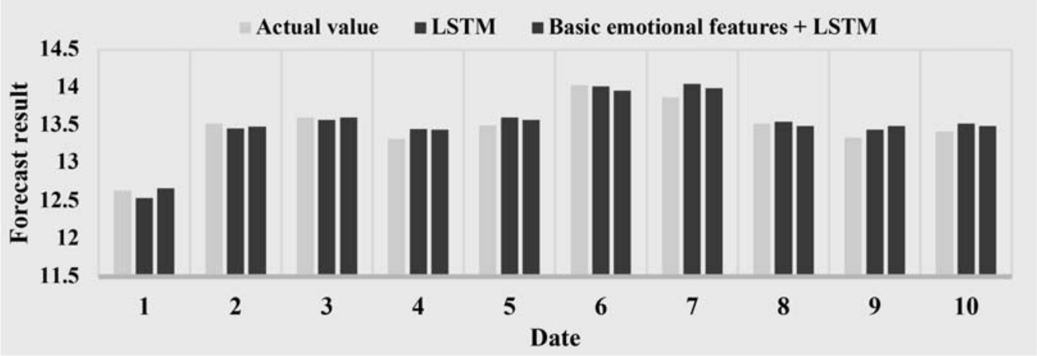

**Figure 4** Short-term stock forecast.

**Table 2** Decomposition of variables on the variance of stock price volatility.

| Period | S.E. | P | RF | SMI | HML | PE | NE | RE | CO |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.3739 | 100.00 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 2 | 0.6379 | 73.206 | 0.2015 | 7.5815 | 12.946 | 3.9510 | 1.2217 | 0.0584 | 0.8345 |
| 3 | 0.8854 | 59.800 | 0.8871 | 1.4949 | 21.343 | 2.2331 | 0.6365 | 0.0294 | 0.5199 |
| 4 | 1.0337 | 52.270 | 0.6912 | 16.831 | 26.037 | 3.1289 | 0.4690 | 0.1883 | 0.3825 |
| 5 | 1.1292 | 48.959 | 0.7807 | 14.991 | 31.461 | 2.7092 | 0.3948 | 0.1641 | 0.5383 |
| 6 | 1.1659 | 43.842 | 1.5426 | 16.184 | 30.850 | 2.5414 | 0.9603 | 0.3237 | 0.7549 |
| 7 | 1.1783 | 45.913 | 2.0247 | 16.217 | 30.867 | 2.4896 | 1.1535 | 0.5503 | 0.7832 |
| 8 | 1.2460 | 44.340 | 1.9976 | 16.489 | 31.206 | 2.7752 | 1.0379 | 1.3335 | 0.7578 |
| 9 | 1.2925 | 43.770 | 2.2936 | 15.323 | 32.883 | 2.5978 | 0.9755 | 1.4501 | 0.7054 |
| 10 | 1.3410 | 42.137 | 2.4869 | 14.457 | 35.449 | 2.4162 | 0.9074 | 1.4626 | 0.6824 |

more than irrational sentiments, which reflects the difference between individual and institutional investors.

A comparison between the predicted results of the stock prediction model and the actual value is shown in Figure 3. From the figure, it can be seen that the predicted value of the multi-data source stock prediction model containing the simple sentiment index is closer to the actual closing price. The relative error between the predicted value and the true value does not exceed 1.9% and the prediction trend accuracy rate is 78.9 %. The relative error between the predicted value and the true value of the multi-data source stock prediction model without the simple sentiment index is 6.3%, and the trend accuracy rate is 58.4%.

The short-term stock forecasts of the LSTM model and the LSTM model fused with emotion are shown in Figure 4. This article compares the two LSTM-based stock market prediction models with a number of neurons of 9 with the real value. It can be seen that the change trend of the three curves is basically consistent with the change trend of the real value of the stock, which can intuitively reflect the prediction effectiveness.

The variance decomposition of variables to stock price volatility is shown in Table 2. It can be clearly seen that the changes in the influence of sentiment variables on stock prices first rise and then fall, while the changes in the variables in the three-factor model have an impact on stock prices from small to large. With the passage of time and the digestion of public opinion hotspots, the influence of public opinion information will gradually reduce and the influence of factors in the three-factor model on stock prices will gradually increase.

The model prediction loss function with different sentiment variables is shown in Figure 5. The LSTM model with the "fear" emotion has the best prediction performance under the four loss functions, so overall, the prediction effect of the "fear" emotion is the best. However, the LSTM model without emotion variables ranks lower under multiple loss functions,

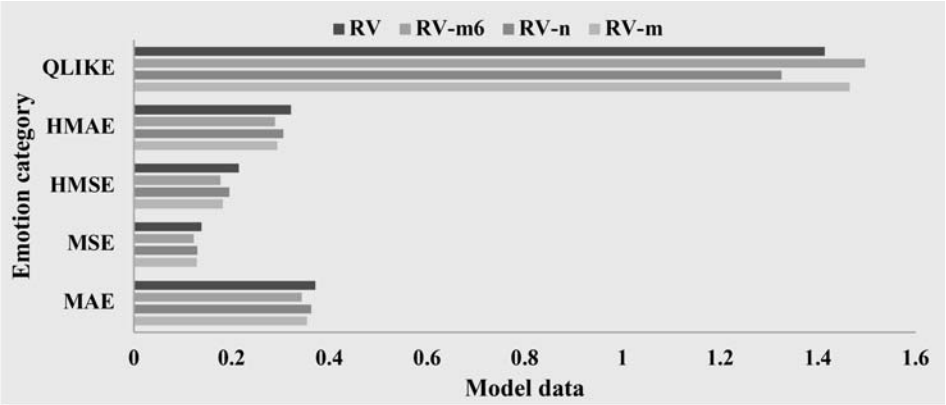

**Figure 5** Model prediction loss function with different sentiment variables.

**Table 3** Classification evaluation results.

| | Recall rate | Precision | F statistics |
|---|---|---|---|
| First step | 0.761 | 0.788 | 0.712 |
| Second step | 0.722 | 0.701 | 0.68 |

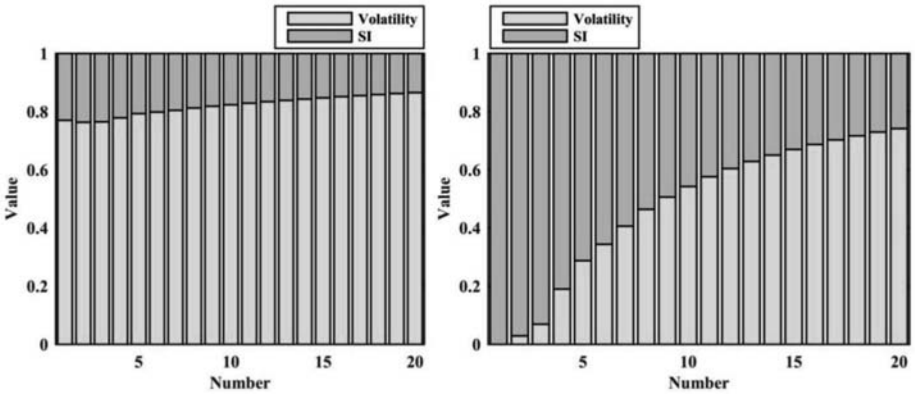

**Figure 6** Variance decomposition running results.

and its overall performance is the worst. The overall predictive performance of negative emotions and Weibo comprehensive emotions is in the middle position.

In the stock market, investor sentiment and stock returns also influence each other as the test results show. When investors are more active in the market, investors have high expectations of the stock market, and the trading volume increases accordingly. This promotes an increase in stock prices and an increase in stock returns. On the contrary, when investors' performance in the market is relatively negative, investors have low expectations of the stock market, selling a large number of stocks, so the stock price continues to fall and stock returns are reduced. However, when stock yields increase, investors are more willing to invest and are more active. On the contrary, when stock returns decrease, investors gradually lose confidence in the stock market and they tend to be hesitant to increase their investment or even sell a large number of stocks. Investors behave more negatively. Therefore, investor sentiment and stock return rate have a mutual influence, and the test results show that they are mutually causal. The classification evaluation results in Table 3 show that of the online stock review data for the 27959 trading days, a total of 23577 related data were obtained, accounting for 84.3% of the total, of which 2,222 were bullish and 1845 were bearish. It can be seen that in the stock forum, most posts provide information related to the stock market, but only a very small proportion (17.2%) of information clearly expresses a view on the rise and fall of the stock market. The vast majority of posts provide objective information related to the stock market, indicating that the stock forum can provide meaningful information to assist investors' decision-making.

Impulse response analysis is used to examine the dynamic reaction process between investor sentiment and the volatility of the Shanghai Stock Exchange Index.The running results of the variance decomposition of $R$ in Figure 6 shows that, the Shanghai Composite Index has a significant volatility cluster, and the contribution of investor sentiment to the volatility of the Shanghai Composite Index return is only 13%. In the long run, the influence of investor sentiment on changes to the Shanghai Composite Index remains basically stable. This result shows that the volatility of the Shanghai Stock Exchange Index is an important factor that cannot be ignored in long-term changes in investor sentiment.

As reported in this article, the stock data of 97 listed companies from March to July 2016 is selected. Growth capacity and profitability are the main factors reflecting the

**Table 4** Stock data.

| Stkcd | Clpr | EPS | AccumFundPS | Income PS |
|---|---|---|---|---|
| 002758 | 81.85 | 0.56 | 5.49 | 16.14 |
| 002759 | 36.18 | 0.38 | 2.93 | 2.9 |
| 002763 | 35.6 | 0.78 | 3.93 | 6.4 |
| 002773 | 66.12 | 0.56 | 1.96 | 3.27 |
| 002775 | 45 | 0.57 | 4.99 | 6.23 |
| 002792 | 43.96 | 0.59 | 1.83 | 3.45 |
| 300008 | 24.92 | 0.22 | 2.43 | 3.66 |

**Table 5** Stock standardized data.

| Stkcd | Clpr | EPS | AccumFundPS | Income PS |
|---|---|---|---|---|
| 002758 | 2.03 | 0.38834 | 1.39964 | 1.97338 |
| 002759 | 0.742 | – 0.02127 | 0.28718 | – 0.1271 |
| 002763 | 0.19 | 0.88898 | 0.72174 | 0.4281 |
| 002773 | 0.71 | 0.38834 | – 0.1343 | – 0.0684 |
| 002775 | 0.16 | 0.411 | 1.18236 | 0.4011 |
| 002792 | 0.725 | 0.45661 | 0.19083 | – 0.0398 |
| 300008 | 1.40 | – 0.38536 | – 0.1604 | 0.1632 |

**Table 6** Factor analysis results.

| Kaiser-Meyer-Olkin measure of sampling adequacy | | .700 |
|---|---|---|
| Bartlett's sphericity test | Approximate chi-square | 762.317 |
| | df | 36 |
| | Sig. | .000 |

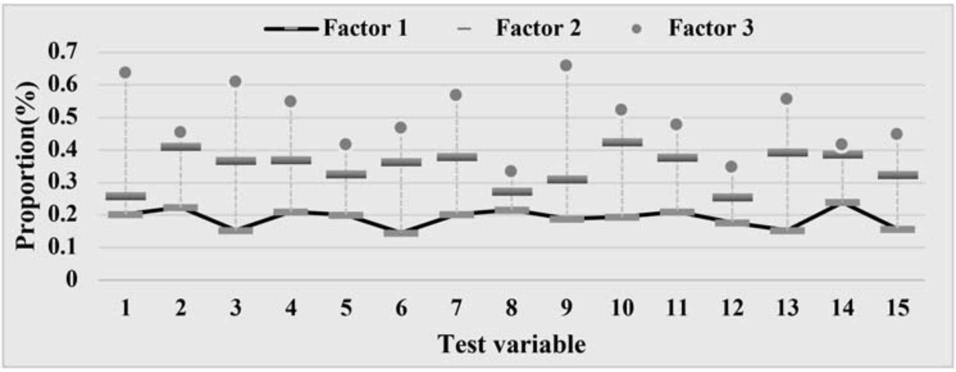

**Figure 7** Rotation component matrix.

long-term development of stocks. nine indicator variables are selected as the criteria for evaluating the stock market: closing price, trading volume, earnings per share (yuan/share), return on net assets, provident fund per share (yuan/share), operating profit per share (yuan/share), net assets per share (RMB/share), operating income per share (RMB/share), and net cash flow from operating activities per share (RMB/share), as shown in Table 4. The text sentiment analysis based on factor analysis classifies the stock market and provides a basis for the analysis and selection of stocks.

The z-score is calculated to standardize the original data as shown in Table 5. Investor sentiment will be affected by interest rates and this effect is negative because the rise in interest rates will increase the cost of capital, and the fundamentals of the country's economy will deteriorate, resulting in a decline in stock market sentiment, and personal and institutional investor sentiment is not high. For the sentiment of institutional investors, the coefficient of the new money supply is positive, which is also consistent with reality. With more money, more funds can flow into the stock market, and the more funds that are supported, the greater the optimism of investors. Similarly, the amount of new currency will positively affect individual investor sentiment.

This study uses KMO statistics to verify the appropriateness of the factor analysis of the eight selected indicator variables. The results are shown in Table 6. It can be seen from Table 6 that the KOM statistic is 0.7 (greater than 0.5), Bartlett's test of sphericity approximate chi-square value is very large, and the significance level is 0.000 (less than 0.01). As there is a correlation between the original indicator variables, the data in this article is suitable for factor analysis.

The orthogonal rotation method standardized by Kaiser is used to rotate the component matrix. The rotated component matrix is shown in Figure 7. It can be seen from Figure 7

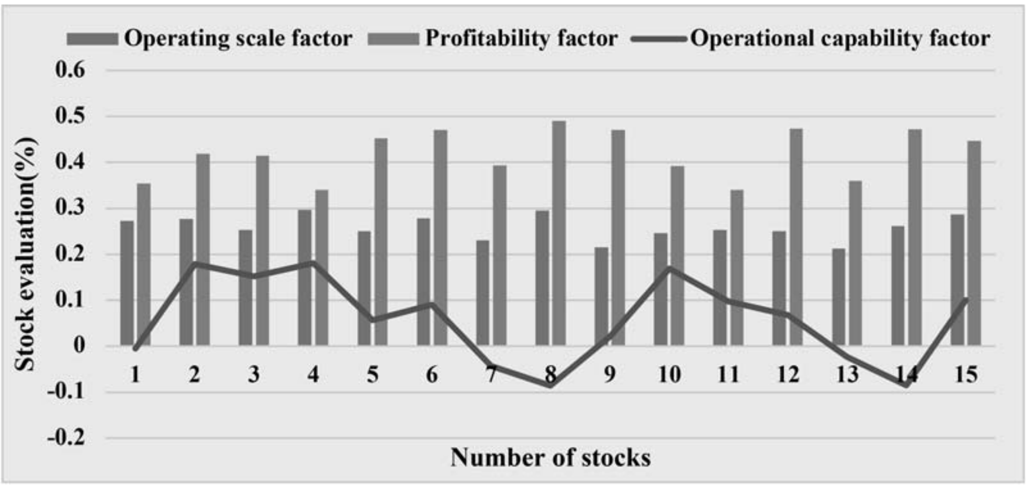

**Figure 8** Evaluation results.

that factor 1 has a greater load on provident fund per share, net assets per share and operating income per share, that is, factor 1 has a greater impact on these three variables and has a strong explanatory effect. Factor 1 is named the operating scale factor. Factor 2 has a larger load on the return on equity, earnings per share, and operating profit per share. Factor 2 can be named the profitability factor. Factor 3 has a large load on trading volume, closing price, and net cash flow from operating activities per share (yuan/share). Factor 3 can be named as the operating capability factor.

The evaluation results are shown in Figure 8. It can be seen from Figure 8 that there are 40 listed companies with a comprehensive score greater than 0.57. The comprehensive score of the company is less than 0.The ranking of individual factor scores will be different from the ranking of the comprehensive scores. Therefore, when evaluating the stock situation of listed companies, it is necessary to combine the three factors to obtain a comprehensive and effective evaluation result as far as possible.

## 5. CONCLUSIONS

This article classifies the sentiment of stock forum comments and calculates the investor sentiment index. Based on the emotional dictionary, the stock market emotional dictionary is formed by adding the professional vocabulary of the stock market. Based on the naive Bayes algorithm, the sentiment classification model is constructed to classify the sentiment of the stock forum comments. There are many factors that affect price trends in the stock market, which contain a lot of noise and are very random. Forecasting the stock market has always been an extremely difficult task. This article provides an analysis from the two perspectives of volume and price information and news text information and combines the two kinds of information through fusion prediction to further improve the prediction effect of the model to obtain greater excess income. The results of the statistical analysis and model prediction show that there is a significant correlation between Internet investor sentiment that reflects investor expectations and stock returns. It can be inferred that the behavior and attitude of investors on online media will have an impact on financial activities, and to a certain extent, provide support for behavioral finance views.

## ACKNOWLEDGEMENTS

This work was supported by Youth Fund Project of Xi’an Traffic Engineering Institute in 2020(20KY-49).

## REFERENCES

1. Huang F, Zhang X, Zhao Z, et al. Image-text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems, 167(MAR.1) (2019):26–37.
2. Singh S K, Sachan M. SentiVerb system: classification of social media text using sentiment analysis. Multimedia Tools and Applications, 78(22) (2019):32109–32136.
3. Xu F, Zhang X, Xin Z, et al. Investigation on the Chinese Text Sentiment Analysis Based on Convolutional Neural Networks in Deep Learning. Computers, Materials & Continua, 58(3) (2019):697–709.
4. Yu C H, Jannasch-Pennell A, Digangi S. Enhancement of Student Experience Management in Higher Education by Sentiment Analysis and Text Mining. International journal of technology and educational marketing, 8(1) (2018):16–33.
5. Liu B. Text sentiment analysis based on CBOW model and deep learning in big data environment. Journal of ambient intelligence and humanized computing, 11(2) (2020):451–458.
6. F Huang, Wei K, Weng J, et al. Attention-Based Modality-Gated Networks for Image-Text Sentiment Analysis. ACM Transactions on Multimedia Computing Communications and Applications, 16(3) (2020):1–19.
7. M, Lutfullaeva, M, et al. Optimization of Sentiment Analysis Methods for classifying text comments of bank customers - ScienceDirect. IFAC-PapersOnLine, 51(32) (2018):55–60.
8. Islam M R, Zibra N M F. SentiStrength-SE: Exploiting Domain Specificity for Improved Sentiment Analysis in Software Engineering Text. Journal of Systems and Software, 145(NOV.) (2018):125–146.
9. Konate A, Ruiying D U. Sentiment Analysis of Code-Mixed Bambara-French Social Media Text Using Deep Learning

Techniques. Wuhan University Journal of Natural Sciences, 23(003) (2018):237–243.

10. Chintalapudi N, Battineni G, MD Canio, et al. Text mining with sentiment analysis on seafarers' medical documents. International Journal of Information Management, 1(1) (2021):1–9.
11. Sohrabi M K, Hemmatian F. An efficient preprocessing method for supervised sentiment analysis by converting sentences to numerical vectors: a twitter case study. Multimedia Tools and Applications, 78(17) (2019):1–20.
12. Alam S, Yao N. The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. Computational & Mathematical Organization Theory, 25(3) (2019):319–335.
13. Zhou F, Zhou H M, Yang Z, et al. EMD2FNN: A strategy combining empirical mode decomposition and factorization machine based neural network for stock market trend prediction. Expert Systems with Applications, 115(JAN.) (2019):136–151.
14. Moews B, JM Herrmann, Ibikunle G. Lagged correlation-based deep learning for directional trend change prediction in financial time series. Expert Systems with Applications, 120(APR.) (2018):197–206.
15. Senarathne C W, Wei J. The impact of patent citation information flow regarding economic innovation on common stock returns: Volume vs. patent citations. International Journal of Innovation Studies, 2(4) (2018):137–152.
16. Oliveira F, Maia S F, Jesus D, et al. Which information matters to market risk spreading in Brazil? Volatility transmission modelling using MGARCH-BEKK, DCC, t-Copulas. The North American Journal of Economics and Finance, 45(JUL.) (2018):83–100.
17. Li L, Hwang N. Do market participants value earnings management? An analysis using the quantile regression method. Managerial Finance, 45(1) (2019):103–123.
18. Alders P, Schut F T. Trends in ageing and ageing-in-place and the future market for institutional care: scenarios and policy implications. Health Economics Policy & Law, 14(1) (2018):1–18.
19. Faghihi-Nezhad M T, Minaei-Bidgoli B. Prediction of Stock Market Using an Ensemble Learning-based Intelligent Model. Industrial Engineering & Management Systems, 17(3) (2018):479–496.
20. Cotti C, Simon D. THE IMPACT OF STOCK MARKET FLUCTUATIONS ON THE MENTAL AND PHYSICAL WELL-BEING OF CHILDREN. Economic Inquiry, 56(2) (2018):1007–1027.
21. J Luan, Shan W, Wang Y, et al. How easy-to-process information influences consumers over time: Online review vs. brand popularity. Computers in Human Behavior, 97(AUG.) (2019): 193–201.
22. Cao H, Lin T, Li Y, et al. Stock Price Pattern Prediction Based on Complex Network and Machine Learning. Complexity, 2019(10) (2019):1–12.
23. Zhou Z, Lin L, Li S. International stock market contagion: A CEEMDAN wavelet analysis. Economic Modelling, 72(JUN.) (2018):333–352.
24. Wang H, Lu S, Zhao J. Aggregating multiple types of complex data in stock market prediction: A model-independent framework. Knowledge-Based Systems, 164(JAN.15) (2019):193–204.
25. Biswas P K, Chapple E. Corporate governance and stock liquidity: evidence from a speculative market. Accounting Research Journal, 33(2) (2020):323–341.

**Haixiang Li** was born in Xi'an, Shaanxi, P.R. China, in 1986. He received a Master degree from Xi'an Polytechnic University, P.R. China. He is currently studying in the School of Economics and Management, Xi'an University of Technology. His research interests include environmental management, industry and regional economic development.
Email:lihaixiang_1986@163.com

**Weixian Xue** was born in Xi'an, Shaanxi, P.R. China, in 1967. He received a PhD degree from Xi'an Jiaotong University, P.R. China. He currently works in the School of Economics and Management, Xi'an University of Technology, His research interests include international trade, investment and environmental management, network economy and electronic commerce and industry and regional economic development.
Email: wxxue2003@163.com