Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 2427389, 9 pages
https://doi.org/10.1155/2022/2427389

*Research Article*

# Stock Market Prediction Based on Financial News Text Mining and Investor Sentiment Recognition

**Jianxin Bi**

*School of Business, Zhejiang Wanli University, Ningbo 315100, China*

Correspondence should be addressed to Jianxin Bi; bijianxin@zwu.edu.cn

Received 29 July 2022; Revised 14 September 2022; Accepted 17 September 2022; Published 5 October 2022

Academic Editor: Zaoli Yang



The stock market is usually regarded as the bellwether of the economy, which can reflect the economic operation of a country or region. As a significant part of the financial market, the equity market plays a critical role in the financial sector. Whether in academia or investment field, stock market forecasts always excite great interest. Financial news is an important source of information in the financial market, which reflects the mood swings of investors and often goes hand in hand with the market trend. However, due to the unstructured and professional characteristics of financial news, there are challenges in accurately quantifying their emotional tendencies. This research is based on Hidden Markov Model (HMM) to segment financial news text. The recognition and classification of news emotion is carried out by bidirectional long short-term memory (BI-LSTM) algorithm, and long short-term memory(LSTM) model is trained with text emotion index and stock market transaction data to realize the prediction of stock market. The results show that BI-LSTM algorithm performs better than the emotional dictionary algorithm in emotional recognition. And the emotional index of financial news text can enhance the accuracy of stock market prediction to a certain extent. Compared with using stock market technical index and news text vector only, the prediction accuracy can be improved by about 2%.

## 1. Introduction

The stock market is often regarded as a barometer reflecting the national economic situation. The fluctuation of stock price can reflect the capital operation and market views of industries or enterprises in a country or region to a certain degree extent. The stock market is a significant reference for analyzing and predicting the economy [1, 2]. With the improvement of Internet infrastructure and the continuous penetration of Internet services, the number of netizens in China has maintained steady growth. Academia and the secondary market are both full of great interest in stock market prediction. Over the years, financial academics have been exploring the inner workings of the equity market and attempting to predict its fluctuations. The stock price has the distinguishing feature of high random volatility; therefore, stock price forecasting is demanding and has poor predictive accuracy [3, 4]. The past stock price is usually utilized to predict the future price [5, 6]. However, some scholars believe that the past price information in the efficient market is useless in predicting the future price [7, 8]. The efficient market hypothesis believes that in a stock market with sound laws, good functions, high transparency, and full competition, all valuable information has been timely, accurately, and fully reflected in the stock price trend, including the current and future value of enterprises. Unless there is market manipulation, investors cannot obtain excess profits higher than the market average by analyzing past prices. In contrast, the fluctuation of stock prices is a known classification problem. Forecasting stock price fluctuations is more feasible and practical than accurate price forecasting.

China's stock market has a huge number of investors, of which retail investor's account for a relatively high proportion. According to the statistics of the fourth quarter of 2021, among the total market value of China's stock market, institutional investors accounted for about 20.3% of the total shares, and individual investors accounted for 22.6%. Retail investors have a short investment cycle and are prone to heavy irrational

emotions. To a certain extent, they lack the professional financial knowledge and the ability to withstand risks required for investment. They are relatively more concerned about short-term policy changes, and their mentality is not mature enough. Retail investors are vulnerable to various news and are prone to emotional chasing up and down and blindly following the trend. The stock market has a variety of trading restrictions in China, and market news and national policies have a great impact on the stock market, resulting in an increase in the probability of irrational trading [9, 10]. Therefore, the structural characteristics of investors with a large proportion of retail investors, strict trading mechanism restrictions, and large policy influence make China's stock market more vulnerable to emotional irrational trading, showing the characteristics of frequent market fluctuations.

In the era of big data, a large number of diversified information and data are being produced at a high speed. People's lives are increasingly dependent on the Internet, and the communication and discussion between investors have also shifted from offline to online. More and more investors get all kinds of information about the stock market, national industries, and companies through mobile apps and financial web pages and exchange and express their investment views on these platforms [11, 12]. The Internet has gradually become an indispensable information exchange medium for individual investors to discuss and exchange views on the stock market, especially for the new generation of young people who are accustomed to the Internet. Individual investors are constrained by their own energy and information acquisition ability and more susceptible to the impact of online media information, resulting in some irrational investment behavior. At the same time, they also publish information and express views [13, 14]. Therefore, the stock forum and other online platforms are playing an increasingly significant role in influencing individual investors' investment decisions. Investors can acquire timely information through the financial website on the Internet, which can provide investors with references for their investment decisions. Whenever there is financial news released at the level of a country, industry, or company, it will form a spreading effect on the Internet. No matter if the news is positive or negative, these news will invisibly affect investors and further affect their follow-up investment decisions, thus having a certain impact on the stock market.

In recent years, the development of natural language processing technology has made it possible to use a large amount of investor opinion comment information. It helps scholars grasp the massive data as a whole and extract the key information they need. Crawler technology can crawl a large amount of information from the Internet and store it effectively. Text mining technology can effectively extract emotional factors in information by using dictionary method and machine learning method, so as to provide technical support for exploring the impact of investor sentiment on stock returns. In this study, we utilize Crawler technology to quickly obtain a large number of financial news reports and utilize text mining technology to analyze the emotion of financial news texts, so as to achieve the emotional measurement of financial news. In the past, researches on the equity market usually utilize the historical data of the equity market, such as trading volume, turnover rate, opening price, and closing price. Usually time series data are used for regression analysis in previous studies to try to find out the pattern of the stock market, while this paper chooses financial news information for the study [15]. Financial news text belongs to unstructured data, which needs to be processed to extract features before further research. Compared with the traditional linear regression method, this study uses BI-LSTM algorithm for classification, which improves the generalization capability of the model.

Different from the previous method of regression analysis using historical stock data, this paper provides a new perspective for the research of stock market prediction from the perspective of text sentiment analysis and machine learning. This study is divided into five parts according to the needs. The first section expounds the importance and complexity of equity market prediction, the influence of investor sentiment on equity market fluctuation, and the equity market prediction method on the basis of text mining. The second part introduces the research status of equity market prediction based on sentiment mining of financial news text. In the third section, the text of financial news is preprocessed by word segmentation method based on HMM, news sentiment classification is realized by BI-LSTM algorithm, and then news sentiment index is calculated. Finally, the stock market prediction is realized by training LSTM model. Section 4 compares the accuracy of text sentiment recognition methods based on the Chinese financial sentiment dictionary algorithm and BI-LSTM algorithm and verifies the predictive effect of news sentiment index on the stock market. The fifth section summarizes the significance and important value of financial news text mining for stock market prediction.

## 2. Related Work

Along with the advancement of the equity market, more and more investors are discovering that there are abundant investment opportunities in the equity market. Financial news is the first-hand news that public investors can get, investors will make investment decisions based on them. Hence, it is of great practical significance for investors to find a method to predict the impact of news on the stock market. With the progress of various computer technologies, the research on the application of machine learning algorithm and text mining technology in the financial field has increased significantly compared with the past. Firth et al. found that companies with low corporate transparency are more vulnerable to investor sentiment than companies with high corporate transparency. It verifies the importance of corporate transparency in alleviating the impact of investor sentiment on stock prices [16]. Li et al. found that when BW sentiment index is used, the causal relationship between sentiment and stock return only exists in the lower quantile. When the consumer confidence index is used as a proxy, the causal relationship between stock returns and emotions becomes significant [17]. Klemola et al. used Google search volume to measure market attention as emotional information and found that changes in the amount of negative search

words such as "market crash" and "bear market" and changes in the amount of positive search words such as "market rebound" can explain the short-term return of stocks [18]. Duan et al. found that the increase of attention will cause the stock market's return for the day and the next two days to decrease, and the greater the investor differences, the greater the short-term trading volume [19]. Hillert et al. used the affective dictionary method to construct daily affective indicators and used the standard deviation of these affective indicators as the proxy variable of opinion differences. It was found that opinion differences were negatively correlated with the next day's return rate and had a certain predictive ability [20]. Ho and Wang classified the news by quantifying the emotional score of the news and constructed an artificial neural network model for predicting stock price fluctuations with the emotional score of the news. The empirical results prove that the prediction model using the emotional score of the news is better than the random walk model [21]. Gunther and Aurelien studied the relationship between environmental, social, and governance news and the stock market. From the perspective of news emotion, these news are divided into positive news and negative news. The research conclusion shows that when companies face negative news, the stock price will decline by an average of 0.1%. When companies face positive news, the stock price does not change [22]. Rahman et al. have built a model that can predict the stock price based on machine learning algorithm and financial news data. The model can recognize the emotion of financial news, and the model has achieved good prediction results [23]. Mo et al. analyzed tens of millions of news articles, and studied the feedback effect between news emotion and stock market return by calculating the emotional score. The research conclusion shows that news emotion has a negative impact on the return when the stock market lags behind for 5 days, and market return has a positive impact on the news emotion when the stock market lags behind for 1 day [24]. Wu et al. build slang emotional word dictionary slangSD, Twitter, and SMS messages are divided into five categories based on dictionary method, with an accuracy of 84% [25]. Malandri et al. compared the best asset allocation strategies constructed by different machine learning models and found that the long-term short-term memory network is better than multilayer perceptron and random forest, and adding emotional data makes the asset allocation strategy perform better [26]. In the above literature, the emotional indicators are set unilaterally in the research of news emotion, and the comparison between different emotional classification methods has not been achieved. This paper establishes more comprehensive emotional indicator and compares the algorithms based on neural network and emotional dictionary.

## 3. Stock Market Prediction Based on Text Mining and Sentiment Recognition

*3.1. World Segment Based on Hidden Markov Model.* Compared with the word segmentation algorithm based on neural network, HMM has the characteristics of fast computing speed, high flexibility, and high accuracy. Therefore, this paper selects the word segmentation method based on HMM. HMM is one of the classical models in machine learning algorithm, which is based on probability and statistics. The process by which a hidden Markov chain generates an unobservable state sequence is described by HMM, and then a sequence of observations is generated from each state. HMM is widely used in natural language processing, speech recognition, pattern recognition, and other fields. In natural language processing, HMM can be applied in word segmentation, part of speech tagging, syntactic analysis, named entity recognition, and other fields based on word tagging. In this paper, we adopt HMM for word segmentation.

HMM consists of hidden layer and observation sequence layer as shown in Figure 1, $S = \{S_1, S_2, \ldots, S_M\}$ indicates a total of $M$ state sets, $V = \{V_1, V_2, \ldots, V_N\}$ indicates a total of N observation sets, $q_t$ represents the state variable of the system at time $t$. Then the probability of transition between different states can be described as $a_{ij} = P\{q_{t+1} = S_j \mid q_t = S_i\}$. After the state is determined as $S_i$, the probability that the observation is obtained can be expressed as $b_{ij} = \{o_t = v_j \mid q_t = S_i\}$, $o_t$ stands for the observation at time $t$. Generally, a HMM can be recorded as $\lambda = [\pi, \mathrm{A}, \mathrm{B}]$, where $\pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$ is the probability of initial state, $\pi_i$ represents the probability that the initial state is $S_i$, $A = \{a_{ij}\}$ is the state transition probability matrix, and $B = \{b_{ij}\}$ is observation output probability matrix.

Decoding is a prediction problem in the practical application of HMM. The HMM $\lambda = [\pi, \mathrm{A}, \mathrm{B}]$ and observation sequence $V = \{v_1, v_2, \ldots, v_T\} v_1 v_2 v_T$ is known, find the state sequence $S = \{S_1, S_2, \ldots, S_T\}$ which maximizes the conditional probability $P\{S \mid V\}$ for the given observation sequence $V$. That is, for a given observation sequence, find the most probable corresponding state sequence. This problem can be described as follows:

$$S^* = \arg\max P\{S \Delta V, \lambda\}. \tag{1}$$

The most classic algorithm to solve the decoding problem is Viterbi algorithm. Viterbi algorithm is also a typical application of dynamic programming. It is based on such a characteristic of the optimal path: if the optimal path passes through node $S_t$ at time $t$, then the partial path of this path from node $S_t$ to terminal $S_T$ must be optimal for all possible partial paths from $S_t$ to $S_T$. Viterbi variable $\delta_t(i)$ is the maximum probability of all single paths $\{S_1, S_2, \ldots, S_t\}$ with state $S_i$ at time $t$. The variable of the node at time $t$-1of the path with the largest probability among all single paths $\{S_1, S_2, \ldots, S_t\}$ with state $S_i$ at time $t$ is called as $\psi_t(i)$. $\delta_t(i)$ and $\psi_t(i)$ play a great role in the Viterbi algorithm, which can be expressed as follows:

$$\begin{aligned} \delta_t(i) &= \max_{S_1, S_2, \ldots, S_{t-1}} P\{S_t = S_i, S_{t-1}, \ldots, S_1, v_t, \ldots, v_1 \Delta\lambda\}, \\ &\quad i = 1, 2, \ldots, M, \\ \Psi_t(i) &= \arg\max_{j=1,\ldots,M} \left[\delta_{t-1}(i) a_{ji}\right], \quad i = 1, 2, \ldots, M. \end{aligned} \tag{2}$$

The specific process of Viterbi algorithm is as follows:

(1) Initialization.

$$\begin{aligned} \delta_1(i) &= \pi_i b_{iv_1}, i = 1, 2, \ldots, M, \\ \Psi_1(i) &= 0, i = 1, 2, \ldots, M. \end{aligned} \tag{3}$$

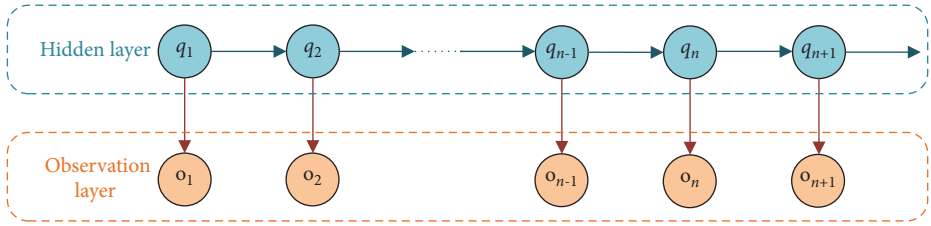


Figure 1: Hidden markov model.

(2) Recurrence. For $t = 2,\ldots,T$,

$$\delta_t(i) = \max_{j=1,\ldots,M}\left[\delta_{t-1}(i)a_{ji}\right]b_{iv_t}, \quad i = 1,2,\ldots,M,$$
$$\Psi_t(i) = \arg\max_{j=1,\ldots,M}\left[\delta_{t-1}(i)a_{ji}\right], \quad i = 1,2,\ldots,M. \tag{4}$$

(3) End.

$$P^* = \max_{j=1,\ldots,M}\delta_T(i),$$
$$S_T^i = \arg\max_{j=1,\ldots,M}\delta_T(i). \tag{5}$$

(4) Optimal path backtracking. For $t = T\text{-}1,T\text{-}2,\ldots,1$,

$$S_t^* = \Psi_{t+1}(S_{t+1}^*). \tag{6}$$

The optimal path is $S^* = \{S_1^*, S_2^*, \ldots, S_n^*\}$. In this paper, the *Python* version of Jieba library is applied for word segmentation. After processing the financial news text, words need to be vectorized. This study uses Word2vec and Google's 13 year open source tool, for vectorization.

*3.2. Text Emotion Classification.* After getting the result of text word segmentation, the emotional tendency of the words contained in each sentence must be classified further. At present, there are usually two methods of emotional classification of texts. One is based on emotional dictionary. Adding vocabulary in the professional field to the existing emotional dictionary through SO-PMI or other methods, and classify emotion and analyze syntactic according to the expanded professional emotional dictionary. It is an unsupervised classification method, which need not to label the corpus. It has an intuitive way of calculating emotional values, and its classification quality is very dependent on the quality of its own emotional dictionary. The other is machine learning algorithm. This method needs the sentences marked with emotion classification as the training set for supervised learning and training model. At present, machine learning algorithms commonly used in building investor sentiment include SVM, Naive Bayes, KNN, and LSTM. To obtain high accuracy, the sentiment dictionary method must consume a lot of labor and time for labeling and requires high quality of labeling, while the BI-LSTM method requires a relatively lower number of samples. The Bi-LSTM method is adopted to classify the emotion behind the text in this paper.

LSTM is an advancement of RNN, which can effectively tackle the long-term dependency problem of RNN. LSTM is suitable for processing sequence data, and its complex structure can avoid the gradient disappearance problem encountered by RNN, thus improving the classification accuracy. LSTM has three well-designed gate structure as shown in Figure 2.

Forgetting gate reads $h_{t\text{-}1}$ and $Xt$, and output a value between 0 and 1 by activating the sigmoid function. 1 means "keep completely," 0 means "discard completely."

$$f_t = \sigma\left(w_f\left[h_{t-1}, X_t\right] + b_f\right), \tag{7}$$

where $w_f$ represents the weight, $b_f$ represents the offset, and $\sigma$ represents the activation function.

Input gate determines what value will be updated. Firstly, the sigmoid activation function determines which part of the content to retain, and then the tanh activation function determines its weight and divides its importance. Finally, the calculation results of forgetting gate and output gate are applied to update the cell state $C_t$.

$$i_t = \sigma\left(w_i\left[h_{t-1}, X_t\right] + b_i\right),$$
$$a_t = \tan h\left(w_a\left[h_{t-1}, X_t\right] + b_a\right), \tag{8}$$
$$C_t = C_{t-1} * f_t + i_t * a_t,$$

where $w_i$ and $w_a$ represents the weight, $b_i$ and $b_a$ represents the offset, and $\sigma$ represents the activation function.

Output gate is utilized to determine the output content, and the sigmoid activation function determines which part of the output to retain, and then the tanh activation function determines its weight and divides its importance.

$$o_t = \sigma\left(w_o\left[h_{t-1}, X_t\right] + b_o\right),$$
$$h_t = o_t * \tan h\left(C_t\right), \tag{9}$$

where $w_o$ represents the weight, $b_o$ represents the offset, and $\sigma$ represents the activation function.

The prognosis of the next time output can be only predicted based on the value of the previous time through RNN and LSTM. However, in some problems, the output of the current time not only has a bearing on the state of the previous time, but also on the state of the future time. To judge the sentiment tendency of a word, it is necessary to not only judge based on the preceding text, but also consider the content of the text that follows it, so that the sentiment can be truly judged based on the context. Therefore, BI-LSTM is proposed to improve the classification accuracy. BI-LSTM is generated by combining the forward LSTM with the backward LSTM as shown in Figure 3.

When the financial news is judged to be positive, the output of BI-LSTM is 1; when the financial news is judged to

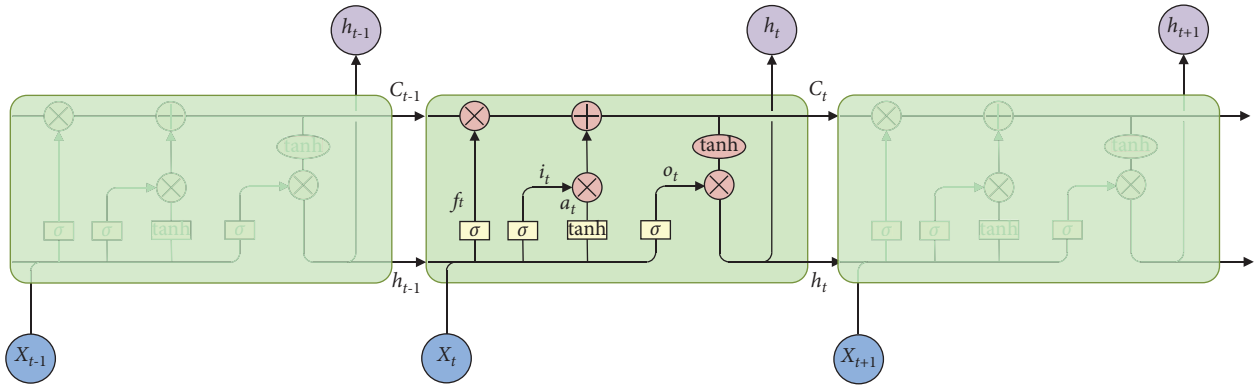


Figure 2: LSTM model.

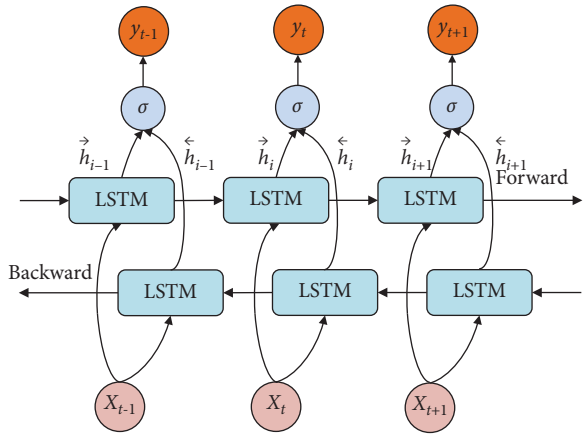


Figure 3: BI-LSTM model.

be negative, the output of BI-LSTM is -1; and when the financial news is judged to be neutral, the output of BI-LSTM is 0.

*3.3. News Sentiment Indicators.* After obtaining BI-LSTM to mark the full sample text information, all emotion classification results are counted and sorted out, and the number of monthly negative emotion posts and positive emotion posts are obtained in this paper. This paper refers to the method of Antweiler and Frank [27] and constructs bullish indicators based on the classification of stock bar posts, namely,

$$\beta_t = \frac{P_t^p - P_t^n}{P_t^p + P_t^n}, \tag{10}$$

where $P_t^p$ represents the number of positive posts at time $t$, and $p_t^n$ represents the number of negative posts at time $t$. The stock evaluation bullish index $\beta_t$ is between −1 and 1, which expresses the relative bullish degree of investors. When all posts at time $t$ are positive, $\beta_t = 1$, which is the maximum value of $\beta_t$. When all posts at time $t$ are negative, $\beta_t = -1$, which is the minimum value of $\beta_t$. This index has nothing to do with the total number of posts.

In addition, Antweiler and Frank also defined another indicator

$$\beta_t^* = \beta_t \ln\left(1 + P_t^{\text{all}}\right), \tag{11}$$

where $P_t^{\text{all}}$ represents the number of all posts at time $t$. This indicator not only considers the relative bullish degree, but also considers the number of posts at time $t$, because the number of posts is also an important standard to measure the strength of investor sentiment.

After considering the number of postings, the "$\beta$" _ "t"^ * indicator does not reflect the consistency of news text sentiment, so a new indicator needs to be constructed to judge the degree of unity of news text sentiment. The convergence index of investors' opinions is adopted to reflect the consensus of different investors in a certain period of time, which is defined as follows:

$$\alpha_t = 1 - \sqrt{1 - \beta_t^2}. \tag{12}$$

The index $\alpha_t$ reflects the degree of disagreement among investors, and the value is between 0 and 1. When investors are unanimously bullish or bearish, $\alpha_t$ takes the maximum value of 1. When investors' opinions are scattered, $\alpha_t$ takes the minimum value of 0.

*3.4. Stock Market Forecast.* This paper generates technical indicators for stock indices based on historical stock trading data. The technical indicators are mainly constructed based on the article of Zhang et al. [28]. The historical trading data of the stock market and the sentiment index of financial news text are used as input data training, and the LSTM model is used to make a more comprehensive prediction of the stock market. As shown in Figure 4, after the emotional indicators of financial news text are extracted, the training results are used as the input of the general model, and the model is finally processed with the full-connection layer, which transforms the stock prediction into a classification problem to predict the future rise and fall.

The output is 1 when the LSTM predicts the stock market to rise, −1 when the LSTM predicts the stock market to fall, and 0 when the LSTM predicts the stock market to be flat.

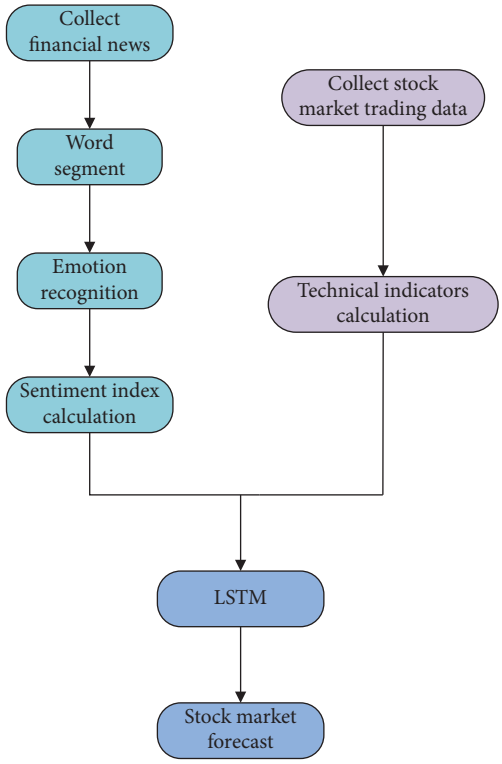


Figure 4: Stock market forecast model.

## 4. Empirical Analysis

*4.1. News Sample Selection and Data Source.* By using selenium tools in *Python* language, this paper collected 95814 pieces of financial news related to CSI 300 constituent stocks companies from January 1, 2020 to May 31, 2022. The representative stock market indexes in the financial market are all from the Wind database. Based on the above information, this paper generates the monthly yield curve of CSI300 stock index, and generates the monthly sentiment index curve of the news based on the emotion dictionary algorithm and BI-LSTM algorithm for the news data set as shown in Figure 5. An emotion dictionary algorithm based on the Chinese LM financial emotion dictionary proposed in [29] is adopted in our work.

The abscissa of the chart represents the time span, that is, from January 2020 to May 2022, a total of 29 months, and the ordinate represents the range of sentiment index and yield. The red line represents CSI300 yield, the blue line represents the news sentiment index constructed by the emotion dictionary algorithm, and the green line represents the news sentiment index $\beta_t^*$ constructed based on the BI-LSTM algorithm used in this paper. As can be seen from Figure 5, compared with the blue line, the trend of the green line and the red line is more consistent, that is, the trend of the news sentiment index generated based on the BI-LSTM algorithm is more consistent with that of the stock index yield curve, and there is a powerful positive correlation between the green line and the red line. It can be seen that the emotional indicators generated based on BI-LSTM algorithm in this paper are better than those generated by the emotional

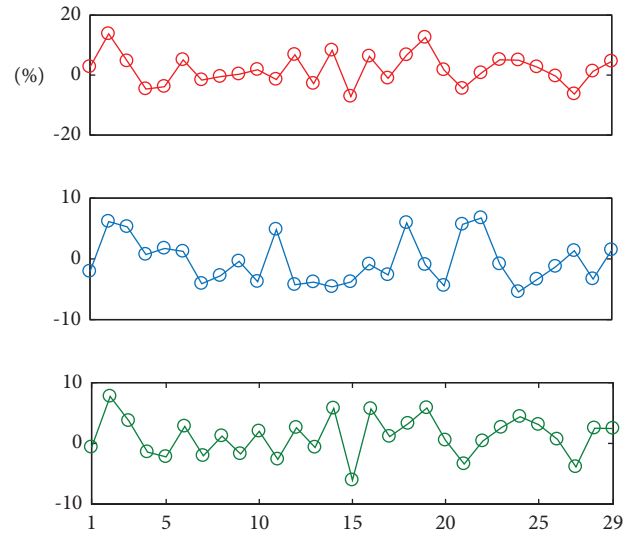


Figure 5: Comparison chart of stock index and news sentiment index.

dictionary method. The reason may be that the emotional index generation method used in this paper better extracts the emotional semantic features related to the fluctuation of stock index in the current news, which makes the calculated news emotional index and stock index trend more consistent.

Next, this paper will further illustrate the consistency between the stock index yield curve and the sentiment index curve by counting the proportion of days of the same rise and fall in the whole time cycle. There are 586 trading days in the news collection cycle. As shown in Figure 6, this paper generates the daily sentiment index of news text based on the emotional dictionary algorithm and BI-LSTM algorithm, and calculates the same rise and fall ratio between the sentiment index and the stock index daily yield.

The abscissa represents different interval days, which are expressed by period. The interval days are 1 day, 5 days, and 10 days, respectively. The blue bar represents the proportional value of the same rise and fall of different stock index returns and sentiment index based on the emotion dictionary algorithm at different time intervals, and the red one based on BI-LSTM algorithm. It can be analyzed from Figure 6 that the news sentiment index generated based on the BI-LSTM algorithm is more relevant to the rise and fall of the stock index, and the proportion of the same fluctuation is usually more than 60%.

*4.2. Stock Market Forecasting.* This paper uses the stock market prediction model proposed in Section 3.4 to realize the short-term prediction of stock index trend, that is, to predict the rise and fall of the next trading day. Each transaction data input into the prediction model includes daily opening price, closing price, maximum price, minimum price and trading volume, technical indicators, and news sentiment index. The data label is obtained by calculating the yield $R$ of the day and converting it through the following formula:

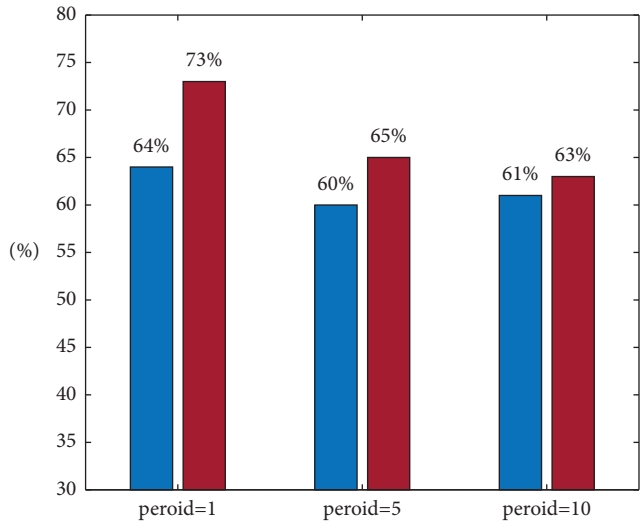


Figure 6: The ratio of sentiment index and stock yield rise and fall at the same time.

$$y = \begin{cases} 1, R > 0.5\%, \\ 0, -0.5\% \le R \le 0.5\%, \\ -1, R \le 0.5\%. \end{cases} \quad (13)$$

The difference between real and predicted stock market yield is shown in Figure 7. Obviously, there is a large deviation between the actual value and the predicted value, but the accuracy in predicting the increase and decrease of stock market yield is more reliable.

In the stock market, there is a common phenomenon of cycle. Time window is an application method of cycle. By setting different time window hyperparameters, the optimal time window can be compared to obtain the best model prediction. In order to select the optimal historical time span of the training data, the historical period $T$ will be added to the model as a hyperparameter. The experimental results are shown in Figure 8.

As shown in Figure 8, the effect of different length time window $T$ on prediction is counted in this experiment, and the time windows of 5, 10, 15, and 25 days are selected, respectively. The blue line represents that technical indicators are adopted only, the red line represents that the technical indicators and news vectors are adopted meanwhile; and the yellow line represents that technical indicators and news sentiment indicators are adopted meanwhile. When $T$ equals 5, the historical data of the first 4 days is used to predict the rise and fall of the stock market on the 5th day. Among them, when $T$ equals 10, the improvement effect is the most obvious. In other time periods, using the financial news sentiment index variable proposed in this paper, the accuracy of the prediction model has been improved to varying degrees. Although adding vectorized original news text to the prediction model also helps to improve the prediction effect, the improvement of accuracy is not as obvious as adding an emotion index, and the complexity of

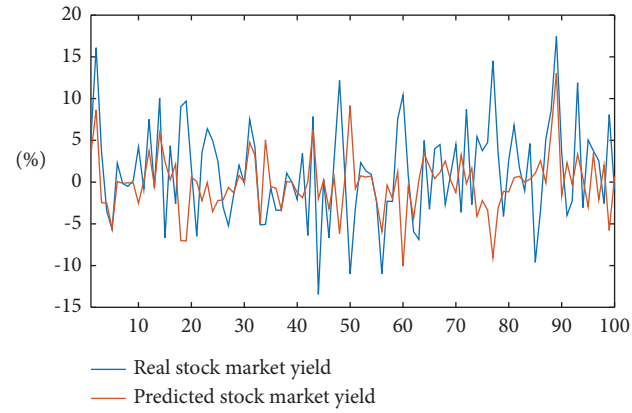


Figure 7: The difference between real and predicted stock market yield.

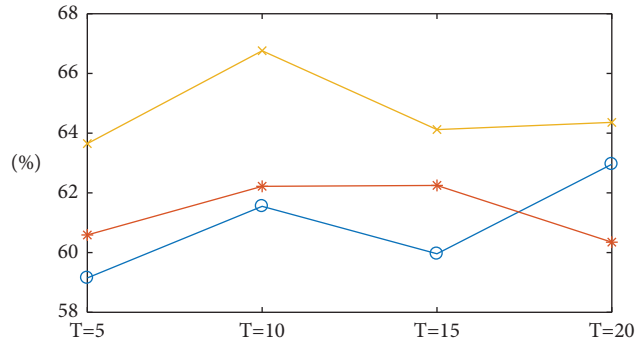


Figure 8: The experimental results with different data feature and historical period.

the model will be greatly increased due to the dimension explosion of the former feature.

Through the double sample $t$-test, we can judge whether the overall mean values of different groups of data in experimental results are equal, and then demonstrate the accuracy of the results from the perspective of statistics. In Table 1, the experimental results in the above table are tested by two sample $t$-test. From the results, it can be analyzed that the feature set composed of news sentiment index and technical indicators is markedly better than other feature sets in terms of prediction accuracy. From the previous empirical analysis results, the emotion index variable can improve the prediction accuracy of the model.

This paper compares the impact of two financial news sentiment analysis methods on the stock market forecast results as shown in Figure 9.

The red bar represents the prognosis model based on the emotion dictionary, and the green one is based on the emotion dictionary. It is evident that the prediction result of BI-LSTM algorithm is better than that of emotional dictionary method because the performance of BI-LSTM algorithm in emotion classification is more accurate.

Table 1: Results of two sample *t*-test.

| | Technical indicators only | Technical indicators + News vector |
|---|---|---|
| Technical indicators + news sentiment indicators | t-value/*p*-value<br>2.91/0.01 * | t-value/*p*-value<br>3.01/0.01 ** |

*Note.* * and ** refer to that it is statistically significant at $\alpha = 0.01, \alpha = 0.05$, respectively.

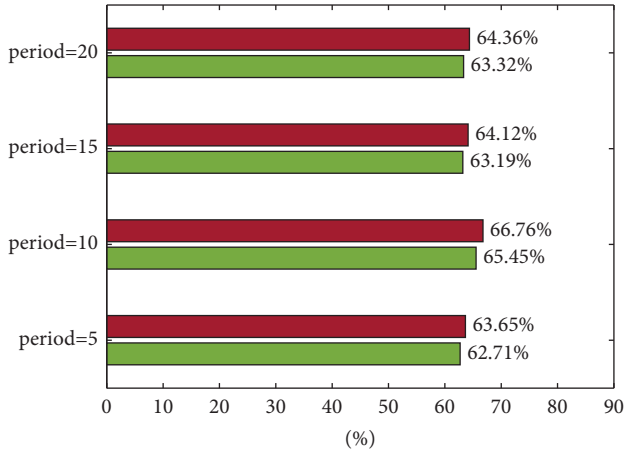


Figure 9: The difference between two emotion recognition methods in stock market prediction.

## 5. Conclusion

Stock is an important means of investment and wealth management, so research on the stock market has important theoretical significance and application value. Since the emergence of quantitative trading in European and American markets, the pursuit of statistical arbitrage to reduce market systemic risks and obtain excess returns has always been the focus of quantitative analysis. With the development of computer technology, quantitative analysis is no longer limited to the field of statistical arbitrage. Text mining technology has been applied to the financial market and has made great achievements.

In this study, financial news text data set is used to generate news text sentiment index, and explore the application of sentiment index in financial market analysis and prediction. Firstly, the financial news text is segmented based on Hidden Markov Model. Considering the influence of context on text emotion, the news emotion recognition is realized through BI-LSTM algorithm, and then the news emotion index is calculated. Finally, the stock market is predicted through LSTM algorithm. In the empirical analysis, the text emotion recognition methods based on the Chinese financial emotion dictionary algorithm and BI-LSTM algorithm are compared. The results show that the emotion classification results based on BI-LSTM algorithm have better consistency with CSI300 component stock return. Then we verify the predictive effect of news sentiment index on the stock market. Taking the calculated financial news text sentiment index, the collected stock market trading data and the constructed technical indicators as the input data set, the LSTM model is trained. Compared with the model that uses technical indicators only, technical indicators and news vectors, the prediction accuracy of the model proposed in this paper can be generally improved by about 2%. Compared with the model of text emotion classification by emotion dictionary method, the prediction accuracy is improved by about 1%. Considering the sentiment indicators of financial news in stock market prediction can improve the accuracy to a certain extent.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the Science and Technology Project of Zhejiang Province (No. 2020C35056); Ningbo Social Science Research Base (Key) Research Project (No. Jd5-zd50); the Basic Scientific Research Funds of Universities and Colleges in Zhejiang Province.

## References

[1] T. Ziarnetzky, L. Monch, and U. Reha, "Simulation-based performance assessment of production planning models with safety stock and forecast evolution in semiconductor wafer fabrication," *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 1, pp. 1–12, 2020.

[2] S. M. Chen and S. W. Chen, "Fuzzy forecasting based on twofactors second-order fuzzy-trend logical relationship groups and the probabilities of trends of fuzzy logical relationships," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 391–403, 2015.

[3] G. Nicolas, M. Andrei, G. Bradley, and A. Liu, "Deep learning from 21-cm tomography of the cosmic dawn and reionization," *Monthly Notices of the Royal Astronomical Society*, vol. 484, no. 1, pp. 282–293, 2019.

[4] Q. Li, Y. Chen, L. L. Jiang, P. Li, and H. Chen, "A tensor-based information framework for predicting the stock market," *ACM Transactions on Information Systems*, vol. 34, no. 2, pp. 1–30, 2016.

[5] C. Qing, J. Ruan, X. Xu, J. Ren, and J. Zabalza, "Spatial-spectral classification of hyperspectral images: a deep learning framework with Markov Random fields based modelling," *IET Image Processing*, vol. 13, no. 2, pp. 235–245, 2019.

[6] S. Li, Z. Zhao, R. Hu, W. Li, and T. Liu, "Analogical reasoning on Chinese morphological and semantic relations," in *Proceedings of the 56th Annual Meeting of the Association for*

*Computational Linguistics*, pp. 138–143, Melbourne, Australia, July 2018.

[7] X. Li, P. Wu, and W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: a case of Hong Kong," *Information Processing & Management*, vol. 57, no. 5, Article ID 102212, 2020.

[8] J. Liew and T. Budavari, "The "sf*a social media factor derived Directly from tweet sentiments*," *Journal of Portfolio Management*, vol. 43, no. 3, pp. 102–111, 2017.

[9] X. Pang, Y. Zhou, P. Wang, W. Lin, and V. Chang, "An innovative neural network approach for stock market prediction," *The Journal of Supercomputing*, vol. 76, no. 3, pp. 2098–2118, 2020.

[10] S. H. I. Yong, Y. Tang, L. Cui, and L. Wen, "A text mining based study of investor sentiment and its influence on stock returns," *Economic Computation & Economic Cybernetics Studies & Research*, vol. 52, no. 1, 2018.

[11] S. Feuerriegel and J. Gordon, "Long-term stock index forecasting based on text mining of regulatory disclosures," *Decision Support Systems*, vol. 112, pp. 88–97, 2018.

[12] L. Chen, W. Su, M. Wu, W. Pedrycz, and K. Hirota, "A fuzzy deep neural network with sparse autoencoder for emotional intention understanding in human-robot interaction," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 7, p. 1, 2020.

[13] A. M. Lozinskaia and V. A. Zhemchuzhnikov, "MICEX index forecasting: the predictive power of neural network modeling and support vextor machine," *Perm University Herald Economy*, vol. 12, no. 1, pp. 49–60, 2017.

[14] M. A. Cheema, Y. Man, and K. R. Szulczyk, "Does investor sentiment predict the near-term returns of the Chinese stock market?" *International Review of Finance*, vol. 20, no. 1, pp. 225–233, 2020.

[15] I. Markvic, J. Stankovic, M. Stojanovic, and M. Stankvic, "Stock market trend prediction using AHP and weighted kernel LS-SVM," *Soft Computing*, vol. 21, no. 18, pp. 5387–5398, 2017.

[16] M. Firth, K. P. Wang, and S. M. Wong, "Corporate transparency and the impact of investor sentiment on stock prices," *Management Science*, vol. 61, no. 7, pp. 1630–1647, 2015.

[17] H. Li, Y. Guo, and S. Y. Park, "Asymmetric relationship between investors' sentiment and stock returns: evidence from a quantile non-causality test," *International Review of Finance*, vol. 17, no. 4, pp. 617–626, 2017.

[18] A. Klemola, J. Nikkinen, and J. Peltomäki, "Changes in investors' market attention and near-term stock market returns," *The Journal of Behavioral Finance*, vol. 17, no. 1, pp. 18–30, 2016.

[19] J. Duan, H. Liu, and J. Zeng, "Analysis on the information content of China's internet stock message boards," *China Academic Journal Electronic Publishing House*, vol. 448, no. 10, pp. 179–192, 2017.

[20] A. Hillert, H. Jacobs, and S. Müller, "Journalist disagreement," *Journal of Financial Markets*, vol. 41, pp. 57–76, 2018.

[21] K.-Y. Ho and W. Wang, "Predicting stock price movements with news entiment: an artificial neural network approach," *Artificial Neural Network odeling*, Springer, Berlin, Germany, pp. 395–403, 2016.

[22] G. Capelle and A. Petit, "Every little helps? ESG news and stock market reaction," *Journal of Business Ethics*, vol. 157, pp. 1–23, 2017.

[23] A. S. A. Rahman, S. Abdul-Rahman, and S. Mutalib, "Minig textual terms for stock market prediction analysis using financial news," in *Proceedings of the International Conference on Soft Computing in Data Science*, Yogyakarta, Indonesia, November 2017.

[24] S. Y. K. Mo, A. Liu, and S. Y. Yang, "News sentiment to market impact and its feedback effect," *Environment Systems and Decisions*, vol. 36, no. 2, pp. 158–166, 2016.

[25] L. Wu, F. Morstatter, and H. Liu, "SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification," *Language Resources and Evaluation*, vol. 52, no. 3, pp. 839–852, 2018.

[26] L. Malandri, F. Z. Xing, C. Orsenigo, C. Vercellis, and E. Cambria, "Public mood–driven asset allocation: the importance of financial sentiment in portfolio management," *Cognitive Computation*, vol. 10, no. 6, pp. 1167–1176, 2018.

[27] W. Antweiler and M. Z. Frank, "Is all that talk just noise? The information content of internet stock message boards," *The Journal of Finance*, vol. 59, no. 3, pp. 1259–1294, 2004.

[28] Y. J. Zhang, L. K. Lei, and Y. Wei, "Forecasting the Chinese stock market volatility with international market volatilities: the role of regime switching," *The North American Journal of Economics and Finance*, vol. 52, 2020.

[29] F. Jiang, L. Meng, and H. Tang, "Media textual sentiment and Chinese stock return predictability," *China Economic Quarterly*, vol. 21, no. 4, pp. 1323–1344, 2021.