Research on Stock Forum Commentary Data Mining and Stock Prediction Based on the LSTM-RF Algorithm

Gang Lei[1], Jinliang Wang[1, 2], Chiyu Shi[3], Junyu Su[3*]
27820999@qq.com, jinliang116@163.com, D22092100293@cityu.mo, D20092100207@cityu.mo

[1]School of Management, Guangdong University of science and technology, Dongguan, China

[2]Faculty of Business, City University of Macau, Macau, China

[3]Faculty of Data Science, City University of Macau, Macau, China

Abstract—As one of the important platforms for investors to exchange their views on stocks, the emotional tendency of stock commentary information influences investors' decisions to a certain extent. The monolithic model is ineffective in analyzing stock investors' sentiment and predicting stock prices. Therefore, we take the Sina stock forum as the entry point and use a web crawler to crawl the comments made by investors of the top 20 stocks in the Shanghai Stock Exchange and use RF to extract and classify the sentiment feature words to build an LSTM-RF stock price prediction model to predict stock price trends. The model can accurately predict the stock price trend.

Keywords- LSTM-RF algorithm; sentiment analysis; stock price prediction

1. Introduction

With the development of the internet, many financial websites today have platforms such as stock forums or forums to give investors a channel of communication and interaction, through which stockholders can express their thoughts freely. As a result, these platform sites encompass a large amount of information on stock commentary.

The stock market in China has developed over 30 years since the establishment of the Shanghai Stock Exchange (November 1990), and although it is not mature compared to mature capital markets abroad, it has been a hot area of research in academia given its ability to help companies raise capital and optimize the allocation of social resources for the development of the Chinese economy. With the development of today's internet and the emergence of social media products such as microblogs, postings, stock forums, and forums, what was once a verbal discussion of the stock market between investors has increasingly moved online, using text-based lines of communication. Not only can this approach be retained and traced, but it can also be captured through crawler technology. To date, there has been a large body of scholarly research into stock commentary on the internet. The findings show that stockholders' investments are often influenced by information on stock reviews on the internet [1]. If there are more positive stock reviews for some stocks, there is a tendency to invest in those stocks. Conversely, if stock reviews are more negative, they are less likely to choose to invest in these stocks. Thus, stock review information on financial websites has a very strong commercial value.

BDEIM 2022, December 02-03, Zhengzhou, People's Republic of China

DOI 10.4108/eai.2-12-2022.2328763

With the passage of time, the number of stockholders' comments on online platforms is growing every moment. Faced with such a huge amount of information, it is certainly difficult to read and analyze it by human operation alone, and computers must be used to do the corresponding work. Sentiment analysis is one of the fastest-growing areas of computer science, and it began in 2003 with a paper by Nasukawa and Yi on product reviews [2]. The goal was to analyze people's emotional dispositions and opinions about entities and their attributes from text, and sentiment analysis has since emerged as a hot topic of interest for scholars and experts. Thomas generates a lexicon of emotive words from comments made by stockholders on the stock market [3]. He suggests that the meaning of a sentence is not only related to the meaning of the words, but also the context and syntax. Therefore, some researchers should analyze and calculate the sentiment tendency of a text about its syntactic structure. To better convey the sentiment in a sentence, the syntactic and structural analysis of the text, together with the inflectional structure, can be used to not only improve the accuracy of the sentiment but also to generate potential targets for the sentiment. Many researchers have used classifiers such as maximum entropy, support vector machines and plain Bayes for text classification [4]. Guo and Sun used R language to crawl the posts of ideal forums, count the total number of posts according to the posting time, calculate the number of posts per day and then estimate the sentiment of stockholders [5]. Yang compared the classification effect of random forest classifier with individual learning and deep learning, and random forest has the advantages of higher accuracy with a low rate and less number of samples required for training [6].

This year, scholars have proposed various models for stock price forecasting. Vargas et al. used deep neural networks to design a stock market trading simulation system for stock market prediction [7]. Wang and Hao used five technical indicators, namely, dissimilar moving average, average trend, relative strength, Bollinger Bands, strong index, and next week's stock price trend as features of the random forest prediction model to construct a GS-RF stock price trend a prediction model based on technical indicators [8]. Liu proposed a stock price prediction method based on the joint application of self-organizing feature mapping (SOM) neural network and a long short-term memory network (LSTM) to help investors screen out stocks with greater profitability [9]. However, the model described above has many drawbacks, such as difficulty in finding the overall optimal solution, low accuracy, slow training speed, etc.

Combining the strengths and weaknesses of the above studies, this paper applies the method of sentiment analysis of shareholder stock reviews taps stock review information and distills the interplay between investor sentiment fluctuations and stock price fluctuations in the Chinese stock market based on the LSTM-RF algorithm to predict stock price directions, which is important for assisting securities investors in making investment choices.

2. Methodology

2.1 RF

A random forest is a classifier that uses multiple decision trees to predict samples. The Random Forest algorithm is a very easy-to-use classification algorithm with stable classification accuracy in most cases.

Figure 1 is the workflow of the random forest algorithm:

(1) First, start with the selection of random samples from a given dataset.

(2) Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

(3) In this step, voting will be performed for every predicted result.

(4) At last, select the most voted prediction result as the final prediction result.

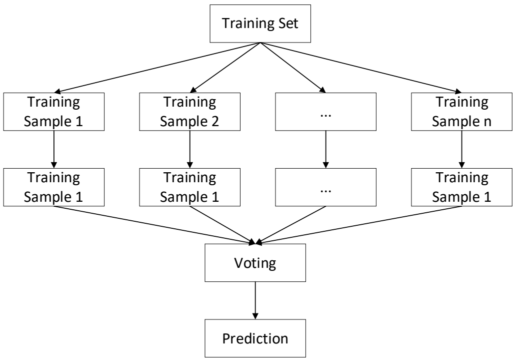

Figure 1. The workflow of the random forest algorithm

2.2 LSTM

LSTM is a special type of RNN, which can be used to control the memory information on the time series by adding memory units, to explore the potential patterns between data and improve the prediction accuracy, and to solve the problem of gradient explosion and gradient disappearance during the training of long series of RNNs.

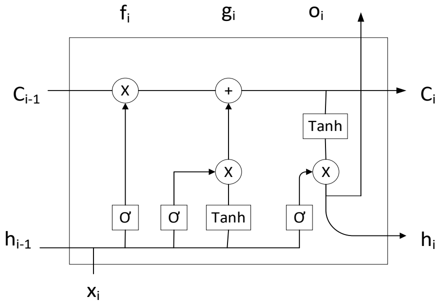

Figure 2. LSTM memory cell

First, through the forgetting gate $f_i$that determines the information that needs to be removed from the neurons,

$$f_i = \sigma(W_f x_i + U_f h_{i-1} + b_f) \tag{1}$$

where σ is the activation function $sigmoid$, sets the information flow weight to a value between 0 and 1, with 0 indicating that the information is removed completely and 1 indicating that all information is retained. $x_i$is the current input vector, and $h_i$is the currently hidden layer vector, andb_f, W_f, U_f are the bias, input weights, and loop weights of the forgetting gate, respectively.

Next, set the external input gateg_i:

$$g_i = \sigma(W_g x_i + U_g h_{i-1} + b_g) \tag{2}$$

Updating neuronal states c_i:

$$c_i = f_i c_{i-1} + g_i tanh(W_c x_i + U_c h_{i-1} + b_c) \tag{3}$$

Output Gates o_i and the information output are controlled by the output gate o_iInformation output controlled by.

$$o_i = \sigma(W_o x_i + U_o h_{i-1} + b_o) \tag{4}$$

$$h_i = o_i \tanh{(c_i)} \tag{5}$$

2.3 LSTM-RF Prediction Model

As shown in Figure 3, to predict stock price movements more accurately, this paper constructs an LSTM-RF-based stock price prediction model with the following steps: `

(1) Data crawling of stock forum evaluation texts with word separation, deactivation, and lexical annotation.

(2) An RF classification model is built to extract sentiment feature words and classify sentiment for investor sentiment.

(3) The LSTM-RF stock price prediction model was constructed; the model was trained and the parameters were adjusted and optimized so that its model loss was less than 0.001.

(4) The model is evaluated using evaluation indicators and finally a forecast of the share price movement is made.

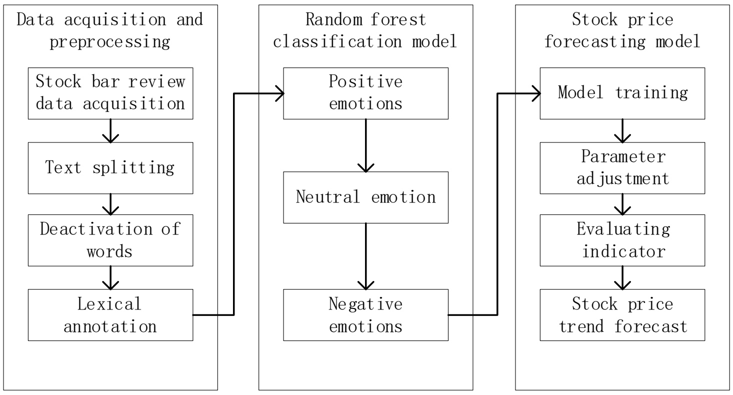
Data acquisition and preprocessing
Stock bar review data acquisition
Text splitting
Deactivation of words
Lexical annotation
Random forest classification model
Positive emotions
Neutral emotion
Negative emotions
Stock price forecasting model
Model training
Parameter adjustment
Evaluating indicator
Stock price trend forecast

Figure 3. LSTM-RF prediction model

3. Experimental Data

3.1 Data Collection

In this paper, we use python to write a crawler and the technical route we take is the Requests library and the Beautiful Soup library. Firstly, the Requests library is used to automatically crawl HTML pages, then the Beautiful Soup library is used to parse the crawled pages, analyze the extraction rules of the data to be collected according to the information organization format of the HTML pages, and finally store the collected data in a structured form [10]. This paper uses a Python web crawler (Requests) to crawl the information of twenty representative stocks in the Shanghai Stock Exchange. The representative Sina stock forum with high user activity and operability was selected as the data source, and the Python crawler was used to crawl pages from the Internet to collect stock comment information from March 31, 2018, to March 31, 2019, and the collected results were stored using text files or structured databases.

3.2 Text Pre-processing

The pre-processing part of this paper includes three parts: Chinese word separation, deactivation processing, and lexical annotation. The general flow chart is shown in the figure 4.

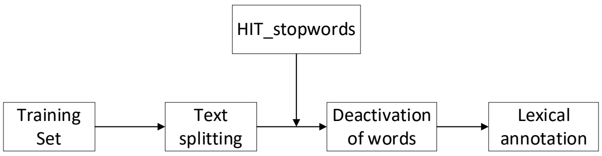

Figure 4. Text pre-processing

Once the data has been collected using the crawler, it needs to be de-noised, split into words, and deactivated. Some of the data were also extracted for sentiment annotation to form a sentiment dictionary. Some of the posts in the stock forum had no body content, and only the post title was used to adopt as content. The crawled text of the stock commentary contains content that is not related to sentiment and needs to be removed.

3.3 Text Splitting

This paper uses the Jieba word-splitting toolkit in python, which is a probabilistic statistical model word-splitting method with relatively stable functions such as word splitting, lexical annotation, and keyword extraction. Jieba has three modes of word splitting, of which the most suitable for text analysis is the exact mode, which allows for precise segmentation of statements.

3.4 Deactivation of Words

Deactivators are special characters and words that have no meaning in the text. For example, words such as "着", "了", "的", and "啊". The text's emotional tendency is not determined by personal pronouns such as "你", "我", "他" and "它". To improve the accuracy of sentiment classification, these noises can be removed during feature traction. In this paper, the deactivated words are removed from the list of deactivated words in HIT, including special symbols, numbers, and functional words.

3.5 Lexical Annotation

Lexical annotation is the process of annotating the lexical properties of segmented words after they have been processed by word separation. There are three criteria for the success of lexical annotation: distributional, morphological and semantic. The lexical annotation is the process of determining the lexical nature of dummy words, adverbs, pronouns, or verbs, etc. Each word is given an exact lexical nature after being processed by the Chinese lexical division technique. There are three main types of lexical annotation methods: random annotation algorithms, rule-based annotation algorithms and hybrid annotation algorithms. In this paper, we use a hybrid annotation method, which is the best annotation method for solving the lexical annotation task.

4. Experimental Results and Analysis

4.1 RF Sentiment Classifier

In this paper, the RF trees are set to 100, 200 and 300, and the feature dimensions of the training data are divided into three groups of 612, 1224 and 2448, using MATLAB's own $TreeBagger$ function. In this paper, 80% of the data are used as the training set and 20% as the test set to select the highest classification accuracy. The best classification result can be obtained when the input feature dimension is 1224 and the number of trees is 200, and the classification accuracy is 86.87%.

This paper compares the classification accuracy (Formula 6) of three different classification methods, RF, KNN, and SVM, as shown in Figure 5. RF has the highest classification accuracy, the best classification effect and can be more accurate for investor sentiment recognition, with an improvement of 13.51% compared to KNN and 5.59% compared to SVM.

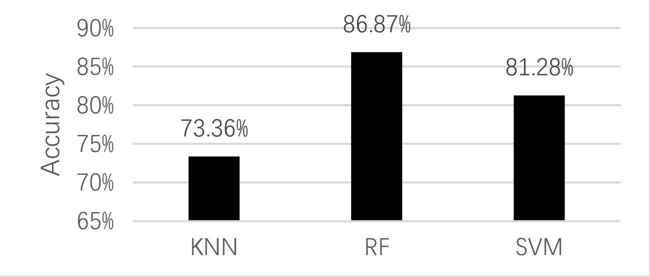

Figure 5. Classification accuracy of different classification models

4.2 Identify the Headings

The LSTM model built in this paper is a GPU version of the Keras framework under the windows operating system, based on the Python language environment. The number of hidden neurons was set to 10, and the first 30 sets of data were used to predict the next set of data, stabilizing the model loss at around 0.001. The Precision, recall, and F1 score were chosen to evaluate the model. The separate random forest model and LSTM model were selected as the comparison models to analyze the price movements of 20 stocks, and the results are shown in Table 1.

The algorithm for the evaluation index is given in the formula below, where TP is true positive, FP is false positive, TN is true negative, FP is false positive.

$$Accuracy = \frac{TP+FN}{TP+FP+TN+FN} \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{9}$$

Table 1 Comparison of Stock Price Prediction Results of Different Prediction Models

Model	***Precision (%)***	***Recall (%)***	***F1- Score***
RF	75.41	78.49	76.91
LSTM	79.77	76.84	78.27
LSTM-RF	84.42	87.65	86.01

As can be seen from Table 1, the combined model used in this paper has substantially improved in terms of accuracy, recall, and model stability compared to the single RF prediction model and LSTM prediction model. Compared to the LSTM prediction model, the accuracy increased by 4.65%, and the recall increased by 10.81%. LSTM-RF can more accurately predict the direction of stock prices based on investor sentiment.

5. Summary

Stock forum comments in the information age are the most convenient channel for investors to express their views on stocks and the fastest and most effective channel for investors to obtain stock information. How to use investor sentiment to predict stock price movements is now a hot topic of research. In this paper, we use the RF classification model to extract sentiment feature words to classify investor sentiment. Comparing the classification accuracy of RF with KNN and SVM classification methods, RF has the highest classification accuracy and the best classification effect, which can be more accurate for investor sentiment identification. A combined LSTM-RF stock price trend prediction model is proposed, which has substantially improved the prediction accuracy, recall, and model stability compared with single algorithm prediction models, and can effectively assist investors in making investment decisions.

Acknowledgment

2021 School level Key Project of Guangdong University of Science and Technology (GKY-2021KYZDW-2)

References

[1] M. Firth, K. Wang, and SML. Wong, “Corporate transparency and the impact of investor sentiment on stock prices,” Management Science, 2015, pp. 1630-1647.

[2] T. Nasukawa, J. Yi, “Sentiment analysis: Capturing favorability using natural language processing,” Proceedings of the 2nd international conference on Knowledge capture, 2003.

[3] R. Thomas, “Intraday online investor sentiment and return patterns in the US stock market,” Journal of Banking & Finance 84, 2017, pp. 25-40.

[4] Q. J. Lan, W.X. Liu, W.K. Li, and X.Y. Hu, “Sentiment Analysis of Financial Forum Textual Message,” Data Analysis and Knowledge Discovery, 2016, pp. 64–71.

[5] K. Guo, Y. Sun, and X Qian, “Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market,” Physica A: Statistical Mechanics and its Applications 469, 2017, pp. 390–396.

[6] H. Y. Yang, W. S. Xing, and C. Chen, “Modular multilevel converter submodule open-circuit fault detection method based on random forest binary classifier,” Proceedings of the CSEE, 2022.

[7] M. R. Vargas, BSLP. De Lima, and A. G. Evsukoff, “Deep learning for stock market prediction from financial news articles,"2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA). IEEE, 2017.

[8] H.Y. Wang, Y.T. Hao, “Stock Price Trend Prediction Algorithm Based on Technical Index and Random Forest,” Modern Computer, 2021, pp. 43–47. (Chinese).

[9] Y.X. Liu, “Analysis and forecast of stock price based on LSTM algorithm." 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI). IEEE, 2021.

[10] J.S. Zhao, M.X. Song, X. Gao, Q.M. Zhu, “Research on Text Representation in Natural Language Processing,” Journal of Software, 2022, pp. 102–128.