

The Prediction Method of Stock Index Based on Sentiment Analysis of Investors' Comments

Jiaming Zhao^{1,*}^a and Ke Wen²^b

¹*School of Management and Economics, Beijing Institute of Technology, Liangxiang East Road, Beijing, China*

²*School of Management and Economics, Beijing Institute of Technology, Beijing, China*

Keywords: Sentiment Analysis, Natural Language Processing, Deep Learning, Stock Index Prediction.

Abstract: The research content of this paper is to obtain the sentiment of retail investors through text analysis of their comments, and apply it into stock price prediction. In this paper, a retail investor sentiment analysis model Fin-Bert-TextCNN is proposed to construct the sentiment index, exploring the relationship between retail investor sentiment and the stock market, and predict the rise and fall of A shares. This paper finds that Fin-Bert-TextCNN model has higher prediction accuracy and is more suitable for the analysis of financial commentary texts. The deep learning technology used in this paper can well depict the nonlinear relationship between investor sentiment and stock index, which is helpful to prevent and resolve major financial risks.

1 INTRODUCTION

In modern society, with the development of enterprise financing, the stock exchange market has injected extraordinary strength into social development. However, in the gradual development of the stock market, many problems have also been exposed. Among them, the financial security problems caused by the excessive sentiments of retail investors have attracted extensive attention in the industry. In particular, the financial crisis or the short selling of international capital may greatly affect the confidence of retail investors and then turn into irreversible turbulence in the stock market. Therefore, it is very important to study the impact of retail investor sentiment on stock trading. In some cases, it is even possible to prevent and defuse major financial risks based on the sentiment of retail investors.

In principle, investor sentiment has a great impact on stock trading. The stock price not only depends on the value of the enterprise, but also related to the investor sentiment to some extent. How to predict the behavior of investors through their sentiments has great research value for ensuring the stability of the stock market. Investor sentiment, especially for the secondary stock market, has a large number of retail


investors, whose sources of information are not only news and industry research, but also subject to a large change in short-term stock price rise and fall. Compared with the news text, the comments of individual investors can better represent their sentiments at a certain time.

However, the existing sentiment analysis model is difficult to extract the sentiment of retail investors in stock reviews. Most of the texts used in the pre-training process are not financial related, and it is difficult to analyze the flexible text of comments in the comment area. Therefore, the core of this paper is to mine the comment sentiment of retail investors through model innovation and then forecast the stock index.

2 LITERATURE REVIEW

There is a strong association between the sentiment of retail investors and stock price movements. Cowles is an early example of stock forecasting by analyzing news texts Cowles (Cowles 1933). He classified Peter Hamilton's editorial articles as "bullish", "bearish" and "uncertain", and then used these classifications to predict the future return of Dow Jones Industrial

^a <https://orcid.org/0000-0003-23338-4444>

^b <https://orcid.org/0000-0003-2377-4000>

Average. Lee et al. (2002) used generalized autoregressive conditional heteroscedasticity to verify the significant positive correlation between investor sentiment and stock returns. Dragos and Laura used consumer confidence index to measure investor sentiment and found it was positively correlated with stock returns (Oprea&Brad 2014). Yang et al. (2020) found that news sentiment has a negative impact on credit default swap spreads, and proved that there was a very close relationship between negative news and the 2008 financial crisis.

With the development of financial markets, many scholars have studied the models of stock price forecasting. Time series model, used as a traditional method, has been difficult to achieve satisfactory results due to the strong model assumptions and inflexible parameter settings (Xu & Tian 2021). On the contrary, literature review shows that deep learning can process higher dimensional data and has stronger representation ability. It has made outstanding achievements in different fields such as speech recognition (Chiu et al., 2018), computer vision (He et al., 2016) and natural language processing (Deng and Yang, 2017). Among them, LSTM model is suitable for modeling time series data, and BERT model has great advantages in financial text sentiment analysis, which can reflect the dynamic changes of stock related companies in real time. Therefore, this paper creatively applies BERT model to the sentiment analysis task of financial comments from investors in China, and proposes a stock index prediction method based on LSTM model.

3 SENTIMENT ANALYSIS MODEL

This paper adopts the Fin-Bert-CNN model when analyzing the sentiment of retail investors. The Fin-Bert-CNN model is divided into two parts: Fin-Bert and TextCNN. BERT is a word embedding model, which is responsible for converting investor's sentiment comment texts into 768 dimensional word feature vectors. The CNN layer model will convert comment texts into sentiment scores after pre-training by extracting the features of word vectors.

3.1 Fin-Bert Word Embedding Model

3.1.1 Bert Word Embedding Model

With the development of machine learning, especially deep learning, more choices have been made in the

construction of sentiment models. The first step of these models for analyzing text of sentiments is to convert the text into word vectors, which is called word embedding. Common words embedding models include Word2Vector (Mikolov et al., 2013) and Glove (Pennington et al., 2014).

Bert (Bidirectional Encoder Representation from Transformers) word embedding model is a language representation model based on Transformer mechanism launched by Google in 2018 (Devlin et al., 2019). BERT, known as a pre-trained natural language processing (NLP) model, uses a new MLM natural language representation algorithm to pre-train bidirectional transformers.

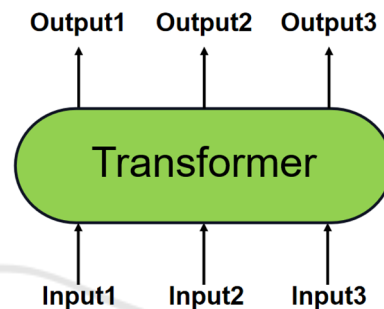


Figure 1: One way Transformer model.

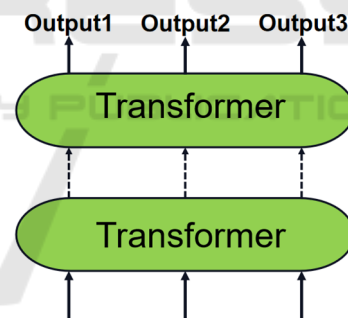


Figure 2: Two way Transformer model.

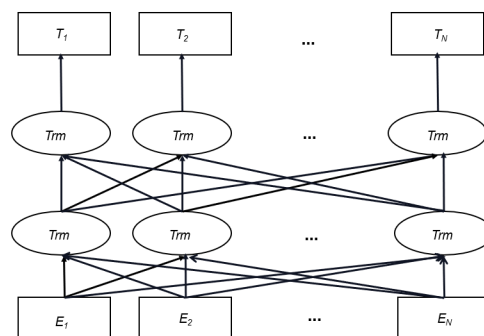


Figure 3: Bert Structure.

As shown in Figure 1 and Figure 2, Bert creatively adopts a two-layer Transformer architecture, and its unique bidirectional coding enables it to depict the representation more accurately. Different from other natural language representation models, Bert does not need a large number of corpora for pre-training. It only needs to add an additional output layer for fine-tune, which can play an outstanding role in various aspects of language processing.

Bert structure is shown in the Figure 3, where E1, E2, ..., En is the text input of the model, T1, T2, ..., Tn is the output vector of the model. The core of bidirectional Transformer feature extractor included in Bert model is self attention mechanism.

3.1.2 Fin-Bert Word Embedding Model Based on Financial Text Pre-Training

The materials used in the traditional BERT model pre-training include the contents of many industries. Although it covers a wide range of fields, it has poor ability to process the financial field corpus. Value simplex re pre-trained BERT prototype by using financial corpora, and made the Fin-Bert word embedding model open-source. Compared with the original Bert-base-Chinese model, the Fin-Bert model is more suitable for financial texts. It can reduce noise and improve accuracy to a certain extent when processing word embedding for retail investor comments.

3.2 Textcnn Convolution Neural Network Model

Convolutional Neural Network (CNN) is a commonly used model in Deep Learning. It includes the input layer, hidden layer and full connection layer (output layer). Multiple neurons in the input layer are responsible for receiving nonlinear input information. In this paper, it is the 768 dimensional word vector output by the Fin-Bert model. The hidden layer is composed of multiple neuron links between the input layer and the output layer. In the hidden layer, the convolution layer and pooling layer appear repeatedly. The final full connection layer DNN is the output layer neuron with the softmax activation function, and the output result is the probability of the possible result, which takes the highest probability as the final output result.

3.3 Steps to Acquire Sentiment Index

We use Fin-Bert-TextCNN model to complete the construction of retail investor sentiment index in the

following four steps. Step 1: Fin-Bert-TextCNN model pre-training. Step 2: Fin-Bert model converts comments into eigenvectors. Step 3: CNN model extracts feature vector characteristics and outputs comment sentiment scores. Step 4: Construct and calculate the sentiment index according to daily retail investor sentiment.

Step 1 input more than 5000 manually scored comments into the sentiment analysis model, and conduct pre-training for the entire model. The Fin-Bert model in step 2 converts the text of investor comments into word vectors. Step 3 is responsible for converting the word vector output from Bert layer into a single sentence score in the same format as the pre-training tag set. For example, 0, 1, 2 represent negative, neutral, and positive respectively.

TextCNN hidden layer includes many alternating convolution layers and pooling layers. Suppose Bert has s words for a single comment. And the word vector output by Bert model has d dimensions. Then for this sentence, we can get the matrix A of row s and column d .

The convolution kernel is a matrix w with a width of d and a height of h . Then w has $h * d$ parameters that need to be updated. For a sentence, matrix A can be obtained after embedding layer. $A [i: j]$ represents lines i to j of A , so the convolution operation can be expressed by the following formula:

The single word window matrix is expressed as:

$$o_i = w \cdot A[i: i + h - 1], i = 1, 2, \dots, s - h + 1 \quad (1)$$

Activate with tan function after stacking offset:

$$c_i = \tan(o_i + b) \quad (2)$$

The convolution result is:

$$c = [c_1, c_2, \dots, c_{s-h+1}] \quad (3)$$

where b represents the offset parameters. Through pooling operation, max pooling is performed on each matrix to obtain a vector z with length m , where m is the number of filters in the convolution process.

$$z = [\hat{c}_1, \hat{c}_2, \dots, \hat{c}_m] \quad (4)$$

The final result is input to the full connection layer and output through the softmax function.

Step 4: Construct the sentiment index according to the daily sentiment of retail investors. Since there are numbers of likes after each comment, it is necessary to calculate the sentiment index through weighted processing.

4 STOCK INDEX PREDICTION MODEL

4.1 LSTM Model

LSTM is an improved RNN model, which has a unique forgetting gate mechanism to solve the problem of gradient disappearance and divergence of RNN in long-term information processing.

LSTM is composed of several consecutive memory cells. The main components of a single memory cell include input gate, output gate and forgetting gate. X_t represents the input at time t , and h_t represents the state of memory cells at time t . σ represents the feedforward network layer of the activation function sigmoid, and \tanh represents the activation function.

In the actual operation process, the input gate inputs data at time t , the input gate and the forgetting gate calculate the data respectively:

$$i_t = \delta(W_i * (X_t, h_{t-1}) + b_i) \quad (5)$$

$$\tilde{C}_t = \tanh(W_c * (X_t, h_{t-1}) + b_c) \quad (6)$$

$$f_t = \delta(W_f * (X_t, h_{t-1}) + b_f) \quad (7)$$

Where, i_t represents the value of the output gate, \tilde{C}_t represents the candidate state of the current cell, f_t represents the value of the forgetting gate, W_o and b_o represents the weight and offset parameters of the forgetting gate respectively. Finally, the status of memory cells is updated and calculates the value of the output gate:

$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1} \quad (8)$$

$$O_t = \delta(W_o * (X_t, h_{t-1}) + b_o) \quad (9)$$

The value of the output gate is output through the activation function tan:

$$h_t = O_t * \tanh(C_t) \quad (10)$$

4.2 LSTM Stock Index Prediction Based on Investor Sentiment

After the Fin-Bert-TextCNN model obtains the sentiment index of retail investors, it is necessary to predict the stock index in combination with other variables. This paper selects six basic characteristics of stock data, namely, the highest price, the lowest price, the opening price, the closing price, the trading volume and the trading price.

The steps of LSTM stock index prediction based on investor sentiments are as follows. First, acquire the basic characteristics of stock data and conduct data preprocessing. Second, input a large amount of stock fundamental data and the corresponding sentiment index into LSTM stock index prediction model for training. Third, predict the stock price for a period of time.

5 RESULTS

5.1 Data Acquisition and Pre-Processing

In this paper, the stock index prediction model is based on the market trading index and investor sentiment index.

Market trading indexes are the six basic characteristics of stock trading data, which are the highest price, lowest price, opening price, closing price, trading volume and trading price. These data can be downloaded from CSI 300 and used after normalization. The method of normalization is selected to use the Min-Max Normalization (Patro et al., 2015), which is also called deviation standardization.

The specific sample data is as follows:

Table 1. Transaction data sample.

Date	Open	High	Low	Close	Volume
2022/3/7	3438.56	3438.56	3360.74	3372.86	3.94
2022/3/8	3372.55	3383.63	3287.34	3293.53	4.16
2022/3/9	3303.71	3321.48	3147.68	3256.39	4.66
2022/3/10	3312.18	3326.58	3291.24	3296.09	3.78
2022/3/11	3259.32	3315.66	3217.42	3309.75	3.84
2022/3/14	3271.89	3297.8	3223.53	3223.53	3.38
2022/3/15	3192.36	3196.92	3063.97	3063.97	4.65

For data representing investor sentiment, this paper uses crawler code to crawl the comment area

from websites such as Eastmoney and Ping An Fortune. This paper has obtained more than 90000

comments, including the date, likes, user ID, and publishing time of each comment. The time span is 44 months.

Due to the obvious noise of comments proposed by retail investors, this paper used the cleaning data algorithm to clear the comments that are blank, or with no actual meaning. Moreover, the disordered symbols in the comments were also removed.

Finally, 92239 qualified comment data were obtained. 4975 of them were selected to be scored manually as the labeled training set of the sentiment analysis model, and 995 of them were used as the test set and verification set.

5.2 Model Pre-Training

The pre-training of the model mainly refers to the investor sentiment model, including fine-tune of Fin-Bert model and CNN model tagged set data. In this part, this paper selected 5000 consecutive data from Eastmoney stock comments for double blind scoring. The scoring standard refers to the sentiment keywords in the industry research papers published by Eastmoney, Wind and other institutions, with the combination of the opinions of experts with years of stock trading experience. Hence, investor comments are divided into three different sentiments with 0, 1, 2 labels representing negative, neutral and positive respectively.

5.3 Results and Verification

In essence, the analysis of investor sentiment is a classification problem. This paper selects the accuracy, precision, recall, and F1 values as evaluation indicators, all of which are commonly used in classification problem. The index calculation method is as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$F_1 = \frac{2 Precision \times Recall}{Precision + Recall} \quad (14)$$

Where, TP represents the numbers of positive and correct samples predicted, TN represents the numbers of negative and correct samples predicted, FP represents the number of positive and wrong samples predicted and FN represents the number of negative and wrong samples predicted. The results are shown in Table 2:

Table 2. Results of sentiment model test set.

Accuracy	Precision	Recall	F1 value
73.17	55.75	55.33	59.97

As it can be seen from Table 2 that the accuracy rate, precision rate, recall rate and F1 value of this model are greatly improved compared with other models, and the accuracy rate of Bert model is more than 70%. At the same time, the results also reversely verifies the theoretical basis of this model, that is, sentiment will have a greater impact on investors' performance and stock index.

The test indicators of stock index prediction are RMSE, SAE and Accuracy. The calculation methods of RMSE and MAE are as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (X_{prediction,t} - X_{real,t})^2} \quad (15)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |(X_{prediction,i} - X_{real,i})| \quad (16)$$

This paper selects three representative market indexes for verification, and the results are shown in Table 3:

Table 3. Prediction results.

Name	Window length	Forecasting time	Accuracy
Shenzhen Stock Index	1017	60	58.90
Shanghai Stock Index	1015	60	60.11
NASDAQ Index	1075	60	78.81

6 CONCLUSIONS

The rise of deep learning technology provides an effective solution for studying the stock financial market and preventing financial security risks. This paper creatively extracts sentiment from the comments of retail investors by constructing Fin-Bert-CNN model and forecasts the stock index through LSTM model.

In the process of exploring the sentiment of retail investors, this paper not only uses the latest pre-trained Fin-Bert model, but also uses the tag set scored under the specialized standard in training process. The results show that this model can extract investor sentiment very well, with an accuracy rate of about 70%. Compared with other natural language

processing algorithms and models, this model is more suitable for the financial industry, especially for retail investors.

Moreover, this paper also plays a positive role in studying the relationship between investor sentiment and stock index. The relationship between investor sentiment and stock index is not a simple linear relationship, but a very complex nonlinear relationship. The deep learning model can well depict this complex relationship. This study also proves that we can predict the rise and fall of stock index with high accuracy after considering the sentiment of investors, especially retail investors. In the future, the study can be improved in the following aspects. Firstly, exclude the survivor bias in investor comments, since profitable investors tend to comment while others not. Secondly, improve the size of data sample, especially the data in tag set, which can make the results more accurate and close to reality.

REFERENCES

- Chiu C.C., Sainath T.N., Wu Y., et al. (2018). State-of-the-art Speech Recognition with Sequence-to-sequence Models, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Cowles A. (1933). Can Stock Market Forecasters Forecast? *Econometrica*, 1(3).
- Deng, L., Yang, L. (2018). *Deep Learning in Natural Language Processing*, Singapore: Springer.
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of NAACL-HLT*.
- He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition*.
- Lee Y.W., Jiang X.C., Indro C.D. (2002). Stock Market Volatility, Excess Return, and the Role of Investor Sentiment. *Journal of Banking & Finance*, 26(12), 2277-2299.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of Advances in Neural Information Processing Systems*.
- Oprea, D.S., Brad, L. (2014). Investor Sentiment and Stock Returns: Evidence from Romania. *International Journal of Academic Research in Accounting Finance and Management Sciences*, 42, 19-25.
- Pennington, J., Socher, R., Manning, C.D., (2014). Glove: Global Vectors for Word Representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.
- S. Gopal Krishna Patro, Kishore Kumar Sahu. (2015). Normalization: A Preprocessing Stage. *CoRR, abs*, 1503.06462.
- Xu, X.C., Tian, K. (2021). A Novel Financial Text Sentiment Analysis-Based Approach for Stock Index Prediction. *The Journal of Quantitative & Technical Economics*, 38(12), 124-145.
- Y, S.X., L, Z.C., Wang, X.J. (2020). News sentiment credit spreads and information asymmetry. *North American Journal of Economics and Finance*, 52.