

# Investor Sentiment Analysis Based on Comment Text for Stock Price Prediction

Zihan Wang \*

Department of Finance, Southern University of Science and Technology, Shenzhen, China

\*Corresponding author: 12010116@mail.sustech.edu.cn

**Abstract.** The sentiment of stock investors is affected by many factors, and the trend of stock price is reflected by the decision behavior of stock trades. Previous tendency can be seen in technical index, and people's sentiment towards the stock is closely related to the text of commentary on the stock. This paper adopts sentiment dictionary to classify the same stock comment text, and modify the corpus of SnowNLP to train a model that conforms to Chinese stock comment. Based on the stock technical index and the established sentiment index to achieve the stock price prediction, so as to achieve a better analysis and prediction of the stock price trend. According to the analysis, the established investor sentiment index has a good effect on stock prediction. To be specific, the predictive regression score of the investor sentiment index is significantly improved compared with those of purely technical indicators, and MAE, MSE and other error indicators are obviously decreased. These results shed light on guiding further exploration of stock forecast with investor sentiment.

**Keywords:** Sentiment; LSTM; RNN; SVM; Stock price prediction.

## 1. Introduction

There has been a debate about whether stock price is predictable or not, and this paper thinks the Chinese stock market is semi-efficient now. Based on Fama's Efficiency Market Theory, the market could not reflect all information [1]. Meanwhile, Nan et al. consider that retail investors account for more than 80% of the trading volume in China's A-share market [2].

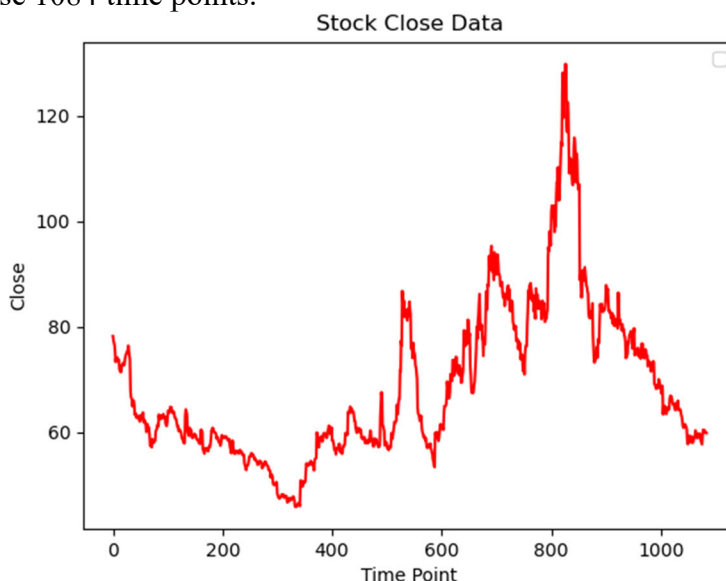
Under such a market environment, this paper considers that stock index prediction is feasible and has certain effects. The prediction of stock based on investor sentiment analysis and technical index is helpful for us to choose stock portfolio in the future. Investor sentiment and technical analysis index will greatly affect the stock price prediction, and the stock price reflects people's inner world of the stock. Wang and Zhao believe that adding a large amount of industry news in stock prediction will reduce the accuracy of prediction [3].

Therefore, this paper chooses a stock belonging to the Sci-Tech innovation board instead of market index for the stock prediction in this paper, which would avoid too much industry news that may affect the accuracy of the prediction of index data such as the market index. The industry news generally has the same effect on the stocks in the same industry, and the selection of individual stocks is also conducive to the analysis of investor sentiment with the help of stocks in the same industry. Scholars have found that when predicting the closing price, every hour investor sentiment has a leading forecasting ability of two hours [4]. This paper took the comments on a stock posted by investors on the Eastmoney Stock Bar of each session to form an analytical indicator of investor sentiment in the first two hours of each session. Neural networks are digital, data-driven, and adaptive [5]. Neural networks have greater advantages in predicting stock volatility than traditional statistical methods. In this training, this paper selected two representative neural network models, LSTM and RNN. Investors usually use enhanced information search to response the uncertain thing and aid decision-making [6]. This study will obtain the comment data about Junshi Bio (688180.sh) from Eastmoney website, analyze the sentiment of investors based on this data based on NLP technology, and combine it with the technical indicators to forecast the stock price. In this way, it provides certain help for stock prediction, and tests whether sentiment analysis indicators based on comment text can help to predict stock prices to provide a basis for subsequent investors to choose stocks.

## 2. Data & Method

### 2.1 Data Acquisition

The Eastmoney Choice financial terminal was used to obtain Junshi Biological 60 minutes line, where the stock data of 1084 hours in 271 trading days is collected from July 5, 2021 to August 12, 2022 in four periods of 10:30, 11:30, 14:00 and 15:00 each day: opening price, high price, low price, closing price, volume. Afterwards, this study uses Java language to climb the comment data in 271 trading days in the Eastmoney Stock Bar respectively. The closing price refers to the price of the last securities traded before the end of trading activity in a day on the stock exchange. Fig. 1 shows the closing price at these 1084 time points.



**Fig. 1** Stock Close Data

The obtained comment section data is shown in Fig. 1. Bullish is positive content, which has the hope that the future stock price increase content. Bearish content is negative content, which has the hope that the future stock price will fall. This study judges the rise or fall of the stock price according to the different attitudes of the shareholders to the future price.

**Table. 1** Some of sample comments.

| Date            | Title                                                                                                                                                  |
|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2022-08-12 1519 | Today came in to bet on the announcement, next week's loss of pants are not left.                                                                      |
| 2022-08-12 1513 | Hold steady, block trading demonstration to retail investors, let retail investors pay chips. Don't be fooled, the COVID-19 drug will be approved soon |
| 2022-08-12 1449 | Here is a joke, The valley is sandwiched between two peaks.-laugh                                                                                      |
| 2022-08-12 1447 | Let's give the idea of operation. Since 524 began to close, it has not stood above 20MA.                                                               |
| 2022-08-12 1427 | Liquor stocks, when to buy are correct, Chinese liquor is always god                                                                                   |
| 2022-08-12 1401 | Daily line KDJ gold fork!                                                                                                                              |
| 2022-08-12 1341 | High short - selling hit dish should be redeemed now. After all, it is enough to have earned at least thirty points or so.                             |
| 2022-08-12 1332 | No one is paying attention to this anymore                                                                                                             |

### 2.2 Data Preprocessing

#### 2.2.1 Data Normalization

In terms of data processing, this paper selected the MinMaxScaler () function to map the feature vectors of stocks to the interval [0, 1]

$$X_{std} = \frac{X - X_{\min}(axis=0)}{X_{\max}(axis=0) - X_{\min}(axis=0)} \quad (1)$$

$$X_{scaled} = X_{std} * (\max - \min) + \min \quad (2)$$

### 2.2.2 Missing value handling

This paper also gives a neutral index of 0 as its sentiment index for the data with no comment data and no sentiment index for that period.

## 2.3 Models

In this paper, three machine learning models, SVM, RNN and LSTM, are selected to judge the changes of stock price prediction compared with technical indicators for the same stock after adding investor sentiment analysis.

### 2.3.1 SVM

SVM is a binary classification model. The basic model of SVM is a linear classifier defined on the feature space which has the largest interval, and the model takes the interval maximization to find the optimal separation hyperplane [7]:

$$w^* * x + b^* = 0 \quad (3)$$

### 2.3.2 RNN

RNN is a recurrent neural network model. RNN solves the problem of processing sequence modeling by introducing hidden state H, which can extract sequential features and take certain transformation as output [8]:

$$O_t = g(V * S_t) \quad (4)$$

$$S_t = f(U * X_t + W * S_{t-1}) \quad (5)$$

### 2.3.3 LSTM

Long short-term memory has a good handle on the problems of gradient disappearance and gradient explosion in the training process of long sequences, it performs better on longer sequences than normal RNN [9]:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (6)$$

### 2.3.4 SnowNLP

At present, there are two methods for Chinese stock comments. First, use the text sentiment dictionary to realize text analysis, annotate the polarity and strength of the sentiment dictionary, and conduct text sentiment classification. Second, one needs to use of machine learning to realize text analysis, label the text information, train the model by supervised learning, and the trained model is used to predict the sentiment polarity of the new text information. In this paper, the two methods are combined. The corpus is trained by text sentiment dictionary and then analyzed by machine learning method. This paper uses the SnowNLP library to retrain the model to help us evaluate investor sentiment:

$$P(c_1|w_1, \dots, w_n) = \frac{1}{1 + \exp[\log(P(w_1, \dots, w_n|c_2) * P(c_2)) - \log(P(w_1, \dots, w_n|c_1) * P(c_1))]} \quad (7)$$

SnowNLP is a corpus-based model that automatically counts word sentiment polarity to judge text sentiment by using corpus of related fields and correlation calculation rules combined with machine learning. The official point out that the high accuracy of e-commerce comments is because its corpus is mainly e-commerce comments data. Therefore, this paper modifies the original corpus and retrain the model to achieve better training results for stock comments.

## 2.4 Metrics

### 2.4.1 MAE

Mean Absolute Error: The mean absolute error is the average value of absolute error, reflecting the actual situation of forecast value error.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (8)$$

#### 2.4.2 MSE

Mean Square Error: The mean square error is generally used to detect the deviation between the predicted value and the true value of the model.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (9)$$

#### 2.4.3 RMSE

Root Mean Square Error: The root of the mean square error represents the sample standard deviation of the difference between the predicted value and the observed value.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (10)$$

#### 2.4.4 $R^2$ Coefficient

$R^2$  Coefficient: The Coefficient of determination reflects how accurately the model fits the data. Generally, the R-squared ranges from 0 to 1. The closer the R Coefficient is to 1, the better the model fit.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (11)$$

### 3. Results & Discussion

#### 3.1 Characteristic Engineering

According to previous studies, investors' sentiment has a great impact on prices within two hours before price formation. Therefore, this paper selects the headline data of comments in the first two hours according to the matching of each hour as the text content of investor sentiment analysis. Due to the emotional analysis of the stock library mostly in English, no convenient processing Chinese stock comments data library of NLP, this paper chose SnowNLP library which is for analysis of buying and selling goods comments sentiment, and updated the trained model of the corpus was, to back up to a certain extent can comment mood emotion of the same kind of stock analysis model. Other stocks in the same industry will also be affected by news in the same industry, and using comments in the same industry as corpus can well avoid personal attitude towards the underlying stock. On the side of updating corpus, this paper based on the same type stock analysis (as in Sci-Tech innovation board or with biological or medicine shares) as the corpus, the method of selecting the Orient Gene Biotech, Fosun Pharma, China-Biotics, Andon health, China Medicine, Tiantan Biological six stocks with each title taken of 10000, a total of 60000 comments to help building a corpus.

Natural Language Processing (NLP) is a subject that uses computer technology to analyze and process natural language. NLP takes computer as a language research tool and can conduct quantitative research on language information under the learning of computer. The paper adopts Machine Intelligence Laboratory of Sichuan University, HIT Database and Baidu database as the JieBa stop words to segment the titles of the 60,000 same industry comments to get the top 1000 most frequent words in the comments. This paper manually signs the 1000 words to determine whether each word is positive, negative or neutral. Then, this paper combines the statistical results and the Chinese financial sentiment dictionary to form a new sentiment dictionary to classify the comment data [10, 11]. This study counts the number of positive words and negative words in the same comment according to the part of speech of the words contained in the sentence after word segmentation in the new sentiment dictionary. Finally, according to the majority of the parts of speech of the positive and negative words to judge the character of the comment title to construct a new positive corpus and negative corpus.

In order to weaken the errors during classification, we combined the new classified corpus with the original corpus of SnowNLP. This paper then trains a new sentiment analysis model using the function of SnowNLP, and then uses the new model to judge the comments selected by the real species and obtains the sentiment index in the range [0, 1]. The index  $>1/2$  is positive content, and the index  $<1/2$  is negative content. In order to construct feature vectors more intuitively, this study performs  $-1/2$  operation on the sentiment index obtained from each comment data to obtain feature quantities belonging to  $[-1/2, 1/2]$ . Therefore, positive and negative comment data are distinguished by positive and negative index. Table. 2 shows partial comments with sentiment index.

**Table 2.** Stock Comment Sentiment

| Title                                                                                                                      | Sentiment |
|----------------------------------------------------------------------------------------------------------------------------|-----------|
| Today came in to bet on the announcement, next week's loss of pants is not left.                                           | -0.49979  |
| Hold steady, block trading demonstration to retail investors, let retail investors pay chips.                              |           |
| Don't be fooled, the COVID-19 drug will be approved soon                                                                   | 0.46099   |
| Here is a joke, The valley is sandwiched between two peaks. -laugh                                                         | 0.499993  |
| Let's give the idea of operation. Since 524 began to close, it has not stood above 20MA.                                   | -0.28411  |
| Liquor stocks, when to buy are correct, Chinese liquor is always god.                                                      | 0.227625  |
| Daily line KDJ gold fork!                                                                                                  | -0.0367   |
| High short - selling hit dish should be redeemed now. After all, it is enough to have earned at least thirty points or so. | -0.42778  |
| No one is paying attention to this anymore                                                                                 | -0.14433  |

Among them, in terms of the emotional factors of each hour, this paper selects to construct the average value of the emotional factors of all the comment data two hours before the time point to weaken the misjudgment behavior in the model and make our processing results present a positive or negative general direction. The Table. 3 lists and summarizes the stock characteristic data of the two days.

**Table 3.** Characteristic Engineering

| Date          | Open  | High  | Low   | Volume    | Sentiment | Close |
|---------------|-------|-------|-------|-----------|-----------|-------|
| 22/8/11 10:30 | 58.22 | 59.73 | 58.01 | 2,553,133 | -0.21089  | 59.4  |
| 22/8/11 11:30 | 59.45 | 61.27 | 59.18 | 3,721,575 | -0.1039   | 60.5  |
| 22/8/11 14:00 | 60.43 | 60.54 | 60.11 | 1,039,927 | -0.08364  | 60.28 |
| 22/8/11 15:00 | 60.28 | 60.54 | 60.15 | 1,148,773 | -0.08281  | 60.5  |
| 22/8/12 10:30 | 60.21 | 60.38 | 59.35 | 1,979,964 | 0.105534  | 60.02 |
| 22/8/12 11:30 | 60.02 | 60.33 | 59.9  | 700,732   | 0.067749  | 60.18 |
| 22/8/12 14:00 | 60.18 | 60.21 | 59.7  | 698,421   | -0.00125  | 59.92 |
| 22/8/12 15:00 | 59.93 | 60.17 | 59.82 | 984,509   | -0.00391  | 59.89 |

**Table 4.** Correlation Data

|           | Open     | High     | Low      | Volume   | Sentiment |
|-----------|----------|----------|----------|----------|-----------|
| Open      | 1        | 0.996677 | 0.99688  | 0.524805 | 0.080987  |
| High      | 0.996677 | 1        | 0.995479 | 0.55857  | 0.094331  |
| Low       | 0.99688  | 0.995479 | 1        | 0.492306 | 0.086708  |
| Volume    | 0.524805 | 0.55857  | 0.492306 | 1        | 0.113643  |
| Sentiment | 0.080987 | 0.094331 | 0.086708 | 0.113643 | 1         |

### 3.2 Correlation Analysis

After the characteristics engineering has been created, this paper carries out the correlation analysis about opening price, the highest and the lowest price, volume, the sentiment index, and finds that the volume and the sentiment index to other three characteristics of the correlation is low and there is a positive correlation relationship. It shows that these two characteristics can be used as predicting factors of stock price (seen from Table. 4).

### 3.3 Model Prediction

This paper used three models, i.e., LSTM, RNN and SVM models, to train and predict the data with and without sentiment index. The results are listed in Table. 5. In the prediction results of the three models, it is not difficult to find that after including the comment emotion factor, the predicted results are better than the data predicted by purely technical indicators in R2 score, RMSE, MSE and RMAE, and the prediction results of SVM model are more obvious. It is obvious that model contains sentiment index prediction has more obvious fluctuation compared with pure technical index, which can better reflect the stock price trend and reflect the stock price trend in the inner world of investors.

**Table 5.** Evaluation of data prediction (Model with \* includes the sentiment)

|         | LSTM     | LSTM*    | RNN      | RNN*     | SVM      | SVM*     |
|---------|----------|----------|----------|----------|----------|----------|
| MAE     | 1.134969 | 0.996933 | 0.792627 | 0.746544 | 4.691244 | 0.806452 |
| MSE     | 2.895706 | 2.745399 | 1.474654 | 1.262673 | 25.42857 | 1.663594 |
| RMSE    | 1.701677 | 1.656924 | 1.214353 | 1.123687 | 5.042675 | 1.289804 |
| R2score | 0.989489 | 0.990034 | 0.980498 | 0.983302 | 0.66372  | 0.978    |

### 4. Limitations & Prospects

There are some limitations in data processing and model selection in this paper. In the part of data processing, it is too slow for the website to crawl the whole review text. Therefore, this paper chooses the "Title" of stock comment for analysis. At the same time, it is difficult to identify whether the comments are published by retail investors or institutional investors in the data crawling process. In terms of model selection, the latest model with attention mechanism is not selected. When manually labeling the characteristic of the most frequently selected 1000 words, it is difficult to ensure that all the labeled data are accurate due to subjective problems. In further research, comment types will be identified and comment sources will be expanded. Meanwhile, when constructing the sentiment dictionary, industry professionals are invited to label the sentiment characteristic of the most frequent words to strengthen the professionalism of classification. In the model, the latest and highly accurate model will also be selected to analyze and forecast the trend of stock prices. Moreover, the influence and positive effect of stock sentiment index on stock price prediction will be further analyzed. The market is changing, the law is changing, history may repeat itself, but it is different, and Jing and other scholars consider that China is a "special" market which subject to trade restrictions and state control [2]. The research and prediction of this paper is also based on the policy implemented in China, a "special" market. Therefore, there are certain regional limitations in the prediction of the trend of stock prices. Future research will be continued to achieve the purpose of analyzing the whole stock market. Hence, it is believed that machine learning still needs some exploration and improvement in predicting the long-term trend based on stock comment sentiment.

### 5. Conclusion

In summary, the experimental results show that incorporating stock sentiment data into the feature quantity is helpful to improve the prediction accuracy of stock prices. In this paper, LSTM, RNN and SVM models are used to predict the stock price of Junshi Biological from the purely technical indicators and the indicators added to the investor sentiment analysis index. Through NLP natural language processing technology and machine learning model to analyze the stock data, and through the continuous adjustment of the model, explore the feasibility of investor sentiment analysis in stock prediction. Among the three models, LSTM model has better prediction performance, and its RMSE finally reaches 1.656924, indicating that LSTM has a good effect in predicting stock price trend, and machine learning can be applied to the stock market and can play a role. There are some limitations in data processing and model selection in this paper, and there are also shortcomings in sentiment analysis in the processing of Chinese information. Nevertheless, this paper finds a method to analyze

the sentiment of Chinese stock comment, which has a certain improvement in stock prediction. In the future, machine learning will be improved enough to predict the long-term stock comment trend. Overall, these results offer a guideline for the stock forecast with investor sentiment.

## References

- [1] Fama E F. The behavior of stock-market prices. *The Journal of Business*, 1965, 38(1): 34-105.
- [2] Jing N, Wu Z, Wang H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, 2021, 178: 115019.
- [3] Wang D, Zhao Y. Using news to predict investor sentiment: Based on SVM model. *Procedia Computer Science*, 2020, 174: 191-199.
- [4] Li Y, Bu H, Li J, et al. The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. *International Journal of Forecasting*, 2020, 36(4): 1541-1562.
- [5] Yu P, Yan X. Stock price prediction based on deep neural networks. *Neural Computing and Applications*, 2020, 32(6): 1609-1628.
- [6] Jiang B, Zhu H, Zhang J, et al. Investor sentiment and stock returns during the COVID-19 pandemic. *Frontiers in Psychology*, 2021, 12.
- [7] Lin Yujian, Liang Ningbo. From unknown to unknown: intelligent security self-evolution. *Communications technology*, 2018, 51(08): 1949-1955.
- [8] Liu Quan, Zhai Jianwei, Zhong Shan, et al. A Deep recurrent Q network model based on visual attention mechanism. *Chinese Journal of Computers*, 2017, 40(6).
- [9] Tian Yi. Research on photovoltaic power prediction technology based on deep learning. Beijing: North China University of Technology, 2019.
- [10] Jiang F, Lee J, Martin X, et al. Manager sentiment and stock returns. *Journal of Financial Economics*, 2019, 132(1): 126-149.
- [11] Jiang Fuwei, Meng Lingchao, Tang Guohao. Media text sentiment and stock return prediction. *Journal of Economics (Quarterly)*, No.4, 2021, pp. 1323-1344.
- [12] nd fuzzy neural network model. *Systems Engineering-Theory and Practice*, 2010, 30(1): 158-160.