Multi-Factor Stock Selection Strategy Based on Network Sentiment Analysis

Daiyu Qian[1,*] and Junru Wang[2]
[1]*School of Economics, Lanzhou University, No. 222 South Tianshui Road, Lanzhou, Gansu Province, China*
[2]*Faculty of Data Science, City University of Macau, Avenida Padre Tomás Pereira Taipa, Macau*
*

Keywords: Network Sentiment, Stock, Regression Analysis, Data Prediction.

Abstract: With the development of the Internet, how people communicate and express their emotions online is more diverse and timely. It has also become easier to access relevant news information in investor networks, and sentiment can play an important role in the stock market, and investors' mood swings can affect the direction of the stock market. The general market sentiment is expressed by panic, greed, and uncertainty among investors. For many individual investors, it is easy to follow the market sentiment and make irrational investments without a professional knowledge background. Based on the previous research results, this paper constructs an investor sentiment indicator by using the method of computer text sentiment analysis and further explores its impact on stock returns. This paper uses the method based on text sentiment analysis to construct the investor sentiment index, and the detailed steps are to first scrape the relevant stock comments from the network directly through the Python crawler, and then use the jieba package to clean the crawled data and extract the weight of keywords. Then, the sentiment index is constructed, which is converted into a sentiment index by collecting the keywords in the comments. Finally, this paper uses the sentiment index as the main factor to calculate the OSL regression to analyze the excess return of stocks.

1 INTRODUCTION

As more social platforms arise, more investors are exchanging ideas and gathering information on them. Researchers can extract valid information from them and use computers to construct algorithms to investigate the influence on investment.

As a more rational investment strategy, quantitative investment minimizes investors' reliance on subjective judgment through mathematical procedures and provides investors with a reference for stock selection when market sentiment swings. The public frequently expects the assets allocated to increase in value. Still, the risk associated with a single asset allocation is significant, so investors build investment portfolios with the hope of avoiding risk while maximizing returns to the greatest extent possible. Computers can analyze and trade quantitative investments, which can improve investment returns to some extent.

Text analysis is a technique for understanding the speaker's emotions and intentions by recognizing and analyzing the text. The dictionary method is the most common: the speaker's conveyed emotions can be identified by matching the keywords to those in the dictionary. Zhao separated text sentiment analysis steps into four categories in 2010: information extraction, information categorization, retrieval and generalization, assessment, and resource development (Yanyan 2010). Wu proposed 2014 the expansion of a common word vocabulary for Web finance based on semantic rule-based text sentiment analysis (Jang 2014). With the continuous development of computer science in China, some common dictionaries continue to emerge. Jiang 2021 pointed out that these dictionaries contain words that are no longer applicable in the financial context, and also omit many financial proprietary sentiment words, which are not applicable in the financial context; it constructed its corpus dictionary based on the financial aspects of the words (Fuwei et al. 2021).

Don't in 2018 proposed that there is a lack of objective evaluation standards in the classification effect, and subsequent research should focus on the construction of sentiment dictionaries in various fields as well as the improvement of sentiment analysis methodology to improve sentiment analysis accuracy (Yalin 2018). In 2018, Liu created an index using sentiment analysis and deep learning

Qian, D. and Wang, J.
Multi-Factor Stock Selection Strategy Based on Network Sentiment Analysis.
DOI: 10.5220/0012829900004547
In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 142-147
ISBN: 978-989-758-690-3

techniques for in-depth mining (Miao 2018). Tian investigated the association between investor sentiment and stock prices in 2019 (Faxiang 2019). Fan explored the heterogeneity of the relationship between investor mood and stock returns of individual stocks in different market capitalization sizes, industries, and market states in 2021 (Pengying 2021). Yang demonstrated in 2022 that the investor sentiment index is better for projecting stocks of small and medium-sized, difficult-to-value companies in the manufacturing industry (Xiaoyu 2022).

It has been noted that the news media adopts shocking terms that may hurt investors. Sentiments generated by social media such as news and stock bar forums can spread rapidly and expand in the market, thus significantly affecting the price formation mechanism (Tetlock et al. 2008). In addition, these sentiments play an important role in the formation of investors' subjective beliefs and have a subtle influence on their perceptions, judgments, and decisions about stock investments (Zhu et al. 2017).

Eastmoney has a leading position in the field of stock trading, its stock bar section has a rich exchange of investor insights, the daily exchange of a huge amount of data, which contains a large number of stockholders on the stock's first emotional reaction, by analyzing their emotions, the paper can achieve part of the stock prediction.

In this paper, we construct the sentiment factor through the methods of text analysis and sentiment analysis, and through the study of the sentiment factor, we observe its effect on the excess return of the stock market. Finally, a regression model is used to compare the prediction effect with and without the sentiment factor, and it is concluded that the prediction model is more stable and accurate when it contains the sentiment factor. The research objective of this study is to provide investors with a more effective basis for decision-making.

2 DATA COLLECTION

2.1 Sample Period Selection

This paper chooses 2022 as the sample, the sample selected from December 31, 2021, to December 31, 2022 trading data. Each stock may have multiple up and down fluctuations during the one-year period, which is reflected in the more obvious stockholder sentiment, and there is a longer period, which is conducive to analysis through market sentiment and has better support for the comprehensiveness of the sample.

2.2 Sample Stock Selection

This paper investigates the top 25 popular stocks in Oriental Finance Network, Oriental Finance Network is the current domestic daily browsing volume at the forefront of the financial website, its popular stock list can reflect the market attention to a certain extent, such as the stock is valued, its liquidity is higher, the market trading is active. It is easy to be influenced by market sentiment, especially in the period of following up and down, the corresponding heat of discussion will be reflected more obviously, so choose the stocks in the hot list for research.

First of all, this paper uses the requests library in Python to crawl the popular page of the Oriental finance network to obtain its top 25 popular stocks and the corresponding code, and then through the Pandas. data frame function to save it to Excel to build the corresponding stock selection form. As shown in Table 1, some representative stocks are listed.

Table 1: Selected stocks.

Name	Code	Name	Code	Name	Code
Yawei Shares	002559	Haima Motor	000572	Embedway	603496
Aotecar	002239	Yinli Media	603598	Sheng-long Shares	603178

2.3 Comment Extraction

In this paper, we use the requests library in Python to capture the discussion posts of corresponding stocks on the Oriental Finance website and use for loop to save the corresponding page number content automatically. Next, regular expressions are used to determine the content of the web page source code to obtain the corresponding title. The posting time range is from December 31, 2021, to December 31, 2022, including the posting time, and title content, collected into Excel waiting for analysis. As in Table 2, some of the comments of the online users of Oriental Finance are listed.

Table 2: Partial display of grabbed comments.

headline	time
It's not good, run!	2022-06-02 11:45:49
Striving to become the world's largest energy storage solution and operation	2022-06-03 16:02:00
The policy is favorable, continue to add to your position! Add positions!	2022-06-02 10:13:47
I'm leaving this afternoon without going up. I don't think it'll work.	2022-06-06 12:08:05
Too trashy. Delisting beats!	2022-06-06 10:54:46
What's wrong with stopping? Just stop so you can add to your position.	2022-06-07 13:13:43

2.4 Segmentation and Getting Emotional Keyword Weights

First, the required libraries were imported, including ‘jilbab’ and ‘pandas’. Then, the required Excel file was read using the `pd.read_excel()` function, and the data was read into the DataFrame object `fm`. Next, concatenate the values of all the cells into a string, using `fm. stack().asype(str)` to convert the DataFrame object into a Series object, and use the `.join()` method to concatenate the values of all the cells and assign the result to the variable `text`. Then, keywords are extracted using the TF-IDF algorithm by calling the `analyze.extract_tags()` function, setting the parameter `text` as the text from which the keywords are to be extracted, `top` as the number of keywords to be extracted, `with eight` as True to indicate that the keywords' weights are returned, and `allows` to specify the allowed of the keywords. Finally, the list of extracted keywords is printed, including keywords and corresponding weights, and the results are stored in the variable `keywords`. As in Table 3, the paper can see the weights covered by verbs, nouns, and adjectives.

Table 3: Crawl keywords and their weights.

verbs
[(‘stop growing’, 0.2627313977640935), (‘stop falling’, 0.14652813031068493), (‘laugh’, 0.1307788377181789), (‘increase a position, 0.12041466058521352), (‘recover’, 0.11256298472109588), (‘launch’, 0.09811677779054795), (‘logout’, 0.0963520622266801 (‘large fall in price’, 0.042674463419117646), (‘rise’, 0.037854672005076556)]
noun
(‘trash', 0.23136287428108518), (‘main force’, 0.1820572323649311), (‘individual’, 0.09325460967583638 (‘stay calm’, 0.08560161030044981), (‘suckers', 0.0845474093346275),
adjectives
[(‘good’, 0.34485513535170276), (‘fantastic’, 0.3220128293469659), (‘advantage’, 0.3013804540119195), (‘rich’, 0.292923361942291), (‘perfect’, 0.06966334557278639)]

Constructing emotional factors:

Setting S^p represents the sum of positive sentiment weights.

$$S^p = \sum_{i=1} Wp \tag{1}$$

Setting S^N represents the sum of negative sentiment weights.

$$S^N = \sum_{i=1} Wn \tag{2}$$

Constructing Overall Sentiment Factors I^T.

$$I^T = \frac{Sp}{Sn} \tag{3}$$

When I^T lies in the interval (0, 1], it represents negative investor performance, and when it lies in (1, $+\infty$], it represents positive investor performance.

After the above methodology, the corresponding investor sentiment factor has been calculated for the selected stocks. As shown in Table 4, the sentiment factor for each stock has been calculated for each month of 2022.

Table 4: XiLong Scientific Monthly Sentiment Factor Display.

Emotional factor	Time
1.020574	2022-1-31
1.089900	2022-2-28
1.429530	2022-3-31
1.182080	2022-4-30
1.214870	2022-5-31
1.190147	2022-6-30
1.209858	2022-7-31
1.124197	2022-8-31
1.246677	2022-9-30
1.240133	2022-10-31
1.206466	2022-11-30
1.262094	2022-12-31

3 BUILD A MODEL

$$\mathrm{Ret}_{t+1} = a_0 + a_1 I^T + a_1 T_t + \varepsilon_t \qquad (4)$$

Where Ret_{t+1} is the stock's excess return in period t +1, a_0 is the intercept of the function, $a_1 I^T$ is the index of the stock's investment sentiment in period t, $a_1 T_t$ is the other chosen indexes in the same period, and ε_t is the perturbation term of the function

In this case, stock excess return, market asset mix factor, book-to-market ratio factor, and market capitalization factor were selected as other factors, so the data corresponding to the time was obtained and the factors were calculated. As shown in Table 5.

Table 5: Selected factors.

Variant	Name	Method
Excess return on equity	Ret	Yield less risk-free rate
Market Portfolio Factor	MP	Total market capitalization-weighted return - risk-free rate
Book-to-market ratio factor	HML	1/Performance Ratio
market capitalization factor	ME	Closing price*Number of shares outstanding
emotional factor	I^T	

3.1 OLS Regression

First, the required libraries were imported, including ‘pandas’, ‘numpy’, ‘matplotlib. pyplot’ and ‘statsmodels. api’. Then, the desired CSV file is read using the ‘pd.read_csv()’ function, specifying the encoding as ‘GBK’. Next, convert the read data to Pandas DataFrame format and assign it to the variable ‘data’. Then, the ‘np. asarray ()’ function was used to convert ‘data’ to the NumPy array format and assign it to the variable ‘np_data’. The data was analyzed with descriptive statistics, including counts, means, standard deviations, etc. Next, specific columns were selected from ‘data’ and they were assigned to the variables ‘x’ and ‘y’. Constant terms were added to the independent variables using the ‘sm. add_constant()’ function and constant terms were added to ‘x’. Then, a least squares regression model was created using the ‘sm.OLS(y,x)’ function, and the model was fitted using the ‘model. fit()’ method. Finally, the ‘model. summary()’ method was used to print detailed summary information about the regression model, including regression coefficients, standard errors, t-values, p-values, and R-squared values.

4 RESULT

The resulting prediction model:

$$\mathrm{Ret}_{t+1} = 0.0207 I^T + 1.9532 MP - 0.2116 HML - 0.0001 ME \qquad (6)$$

Where Ret_{t+1} is the stock's excess return in period t+1, I^T, MP, HML, and ME are the sentiment factor, the market asset mix factor, the book-to-market ratio factor, and the market capitalization factor's factor exposure in period t.

In this paper, a certain number of stocks with the highest excess return in t are continuously screened for buying or selling to maximize the rate of return. Table 6 shows the results of the relevant data obtained after the calculation.

Table 6: Data presentation.

	ratio	Standard deviation	t	P>|t|
emotional index	0.0207	2.018	0.046	<0.001
Market Portfolio Factor	1.9532	2.572	0.011	0.45
Book-to-market ratio factor	-0.2116	-0.636	0.526	-0.87
market capitalization factor	-0.0001	-0.271	0.787	-0.001

For mood index $t < 0.05\, P > |t| < 0.001$ is significant.

For the Market Asset Portfolio factor, although the $t < 0.05\, but\, P > |t| > 0.001$ effects are significant, the hypothesis is more likely to be overturned.

For the book-to-market ratio factor and the market capitalization factor, due to $t > 0.5$, the impact is not significant.

Control group: Selected as regression without sentiment factor

$$Ret_{t+1} = 2.2849MP - 0.2439HML - 0.00003722ME \tag{7}$$

Where Ret_{t+1} is the stock's excess return in period t+1, I^{T}, MP, HML, and ME are the sentiment factor, the market asset mix factor, the book-to-market ratio factor, and the market capitalization factor's factor exposure in period t. As shown in Table7, demonstrates the results of the relevant data obtained after the calculation.

Table 7: Data presentation.

	ratio	Standard deviation	t	$P > \|t\|$
Market Portfolio Factor	2.2849	0.75	3.045	0.003
Book-to-market ratio factor	-0.2439	0.336	-0.725	0.47
market capitalization factor	-0.00003722	0.03	-0.075	0.94

For market portfolio factors $P > |t| < 0.001$ is significant

For the book-to-market ratio factor and the market capitalization factor, the probability of overriding the hypothesis is greater and the impact is smaller due to the greater P>|t|.

5 CONCLUSION

Analyzing the data with and without the sentiment factor, it can be concluded that when containing the sentiment factor, the data performs more stable, especially for the regression of the data on the sentiment factor, and, for the prediction of the accuracy is higher, the probability of overthrowing the established assumptions is minimal, so with the sentiment factor there will be a more perfect prediction model. This study suggests that taking market sentiment into account when selecting stocks for quantitative trading will help to increase the excess return of stocks. Therefore, it is recommended to enhance returns by incorporating market sentiment considerations when making stock picks on the market and constructing multiple portfolios, as well as making timely portfolio adjustments based on market sentiment. This may be because market sentiment is more volatile due to data, and subtle stock price fluctuations will lead to more pronounced market sentiment swings due to investor nervousness. This study enhances stock selection by utilizing the sentiment factor, which is beneficial to extend the study of stock selection factors selection, and allows individual investors' sentiment to be linked to the market, which is informative for other researchers in selecting factors. Finally, this study does not adequately consider the classification of stocks, for different types of stocks their degree of feedback on market sentiment is likely to be different, so it is also necessary to study each type of stock separately, and in the future can be refined on the stock before exploring the relevant conclusions.

AUTHORS CONTRIBUTION

All the authors contributed equally and their names were listed in alphabetical order.

REFERENCES

Z. Yanyan, Q. Bing, · L. Ting. Software Journal. 21.08:1834-1848, (2010).

W. Jang. Computer Applications 34.02:481-485+495, (2014).

J. Fuwei, M. Lingchao, T. Guohao. Economics (quarterly), 21.04:1323-1344, (2021).

B. Yalin. The contemporary economy .06:34-36, (2018).

L. Miaol. Forum on Statistics and Information 33.08(2018):31-38.

T. Faxiang. MA thesis, (2019).

F. Pengying. Mathematical practice and understanding, 51.16:305-320, (2021).

Y. Xiaoyu. MA thesis, (2022).

P. C Tetlock, M Saar-Tsechansky, S Macskassy. Journal of Finance, 63(3):1437-1467, (2008).

Y. Jiaxing, W. Jing. Economic research, 47.07:141-152, (2012).

Y. J. Zhu, Z. Wu, H. Zhang, J. Yu. Chinese Stock Markets, 57:1635-1670, (2017).