Research on Investor Sentiment and Stock Market Performance Based on Text Mining of Online Stock Forums

Sheng Wang [1]

3243138350@qq.com

[1] Dept. of Management, Hubei Business College, Wu Han, Hu Bei Province, China

Abstract: This study examines the relationship between investor sentiment, derived from online stock forum posts on the East Money website, and the closing price and trading volume of the Shanghai Stock Exchange Composite Index (SHCI). The findings reveal that different types of investor sentiment exert distinct effects on the stock market. Negative sentiment, identified through text mining of stock forum posts, demonstrates consistent predictive power for the SHCI's closing price. Moreover, changes in the proportions of positive and neutral sentiment, also extracted from stock forum posts, exhibit an immediate impact on the SHCI's trading volume, with a one-day lag. Thus, the study establishes the significant role of investor sentiment, extracted from stock forum posts, in predicting stock market performance

Keywords: text mining, investor sentiment, stock market performance

1 INTRODUCTION

Asset pricing is a central concern in finance, and different theories offer diverse perspectives. Classical pricing theory emphasizes the role of fundamental information as the primary determinant of asset prices and fluctuations. The efficient market hypothesis posits that if financial markets are efficient in terms of information, asset prices should reflect all available knowledge about market fundamentals. It argues that even if irrational investors exist, their actions will be counteracted by arbitrageurs, resulting in negligible effects on asset prices. However, since the mid-1980s, certain market anomalies have prompted researchers to question the assumption of investor rationality. Several studies have explored alternative possibilities that challenge the validity of the efficient market hypothesis. One such theory is the noise trader theory proposed by Black (1986)[1], DeLong (1990)[2], and others, which suggests that if investors trade based on irrelevant signals, asset prices may deviate from their intrinsic value. This theory assumes two types of traders: rational fundamental traders with unbiased expectations and noise traders relying on different signals with biased expectations. Empirical evidence supports the idea that noise traders can introduce price deviations, and this perspective has garnered increasing acceptance among scholars.

The Chinese stock market exhibits instances of investor irrationality, particularly among retail investors who face information and knowledge limitations, hindering their ability to fully

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334678

rationalize asset valuation (Ba and Zhu 2016[3], Wen et al. 2014[4]). Moreover, imperfect trading mechanisms in the market, such as short-selling restrictions and the immaturity of institutional investors, delay the realization of arbitrage opportunities. Consequently, the pricing of assets in the Chinese securities market warrants investigation, particularly regarding the influence of investor sentiment on asset prices. However, the unobservable nature of investor sentiment poses empirical challenges, prompting the need to determine effective approaches for characterizing investor sentiment. Bu and Pi (2014)[5] examined various methods used in academic literature and identified the synthesis of investor sentiment indices based on market variables proposed by Baker and Wurgler (2006)[6] as the most prevalent approach. Recent research has focused on leveraging internet platforms, including websites, social media, and online message boards, for mining investor sentiment. In the era of big data, exploring ways to effectively utilize abundant internet platform information to extract investor sentiment and understand their trading behavior is a valuable avenue for further investigation.

This study aims to examine the relationship between investor sentiment, as derived from online forum posts, and stock market performance. Specifically, the research investigates whether investor sentiment possesses predictive power for stock market closing prices and trading volume. The Shanghai Composite Index is selected as a representative variable for the overall stock market. To analyze the impact of investor sentiment on the stock market, Granger causality tests are utilized. Empirical analysis is conducted to gain insights into the behaviors of investors engaging in online stock forums discussions and the resulting outcomes. By exploring these aspects, this paper contributes to our understanding of the influence of investor sentiment on stock market dynamics.

2 METHODOLOGY

This study investigates the link between investor sentiment, derived from the text of online forum posts on East Money, and the closing prices and trading volume of the Shanghai Composite Index. To examine the causal relationship between investor sentiment and the stock market's closing prices and trading volume, the Granger causality test is employed.

The specific description of the research steps is as follows:

Step 1: The study utilizes Scrapy, a web crawling framework, to collect all relevant posts from the online stock forum on East Money's website that pertain to the constituents of the Shanghai Composite Index. The collected data is saved as daily text documents, excluding non-textual data such as images, audio, and video.

Step 2: Employing semantic analysis tools, the text of the forum posts is analyzed to determine the sentiment orientation. The study classifies sentiment into three categories: positive sentiment orientation, neutral sentiment, and negative sentiment orientation.

Step 3: Granger causality tests are conducted to examine whether a leading-lag relationship exists between investor sentiment orientation and changes in the closing prices and trading volume of the stock market.

3. RESULT

3.1 Data

The textual data for this study is sourced from the stock forum, a platform hosted by East Money (http://guba.com.cn). A web crawler, specifically Scrapy, is utilized to collect approximately 3 million primary post data from the constituents of the Shanghai Composite Index within the timeframe of January 1, 2022, to April 28, 2023. Additionally, the closing price (CIPr) and trading volume (TradVol) data for the Shanghai Composite Index are selected from the Wind database, covering 330 trading days between January 1, 2022, and April 28, 2023.

The data is standardized using the Z-Score model. The transformation model is as follows:

$$Z = \frac{X - \bar{X}}{\sigma} \quad (1)$$

$\bar{X}$ represents the mean of X, and σ represents the standard deviation of variable X. The processed data conforms to the standard normal distribution with a mean of 0 and a standard deviation of 1.

3.2 Generation of Emotional Time Series

To generate emotional time series, sentiment analysis was conducted on the collected posts using the ROST EA program. This program employs a sentiment lexicon to assign scores to each post based on its tokenized content. A score of 0 indicates a neutral sentiment, while scores above 0 indicate a positive sentiment and scores below 0 indicate a negative sentiment. By aggregating the number of occurrences for each sentiment category within each post, the proportion of positive, neutral, and negative sentiments is determined. The proportion of positive sentiment in the daily posts from the stock forum represents investors' positive sentiment tendency and is referred to as "Positive". Similarly, the measurements for neutral and negative sentiment tendencies among investors are obtained using the same methodology and are respectively labeled as "Neutral" and "Negative".

After applying the ROST EA program to the data, a dataset of 2.48 million valid sentiment tendency data points was obtained for analysis in this study. The analysis of the data revealed that the proportion of positive sentiment in daily posts consistently outweighed the proportions of neutral and negative sentiments. Additionally, a notable observation was made regarding the source of negative sentiment articles, with a majority originating from personal accounts. This trend may be attributed to heightened media scrutiny and regulatory constraints placed on institutional accounts.

3.3 Granger Causality Test

The article conducted the ADF unit root test to examine the stationarity of each time series, and the results are shown in Table 1.

Table 1: Unit Root Test Results for Time Series (Notes: * denote significance for 5%)

	CIPr	Tradvol	Positive	Neutral	Negative
P	0.000*	0.000*	0.117	0.115	0.000*

Table 2: Unit Root Test Results for First-Differenced Time Series. (Notes: * denote significance for 5%)

	DCIPr	DTradvol	DPositive	DNeutral	DNegative
P	0.000*	0.000*	0.000*	0.000*	0.000*

Table 1 presents the results of the stationarity analysis for the closing price, trading volume, and investors' sentiment tendency time series. The p-values for the closing price, trading volume, and negative sentiment tendency indicate stationarity at a 95% confidence level, as they are below 5%. However, the positive and neutral sentiment tendency time series exhibit non-stationarity. To ensure compatibility for the Granger causality test, which requires integrated stationary time series of the same order, all-time series underwent first-order differencing followed by the Augmented Dickey-Fuller (ADF) test.

As shown in Table 2, upon conducting the unit root test on the first-differenced time series, the p-values for the investors' positive sentiment and neutral sentiment sequences are both 0.0000, suggesting that they have achieved stationarity. Consequently, all the time series, namely DCIPr, DTradvol, DPositive, DNeutral, and DNegative, exhibit the same order of stationarity.

To investigate the transmission effect of investor sentiment fluctuations on stock market performance, this study utilizes a Vector Autoregressive (VAR) model, considering the autocorrelation present in stock market closing prices and trading volumes. The VAR models are constructed for six pairs of time series, namely the first-order differenced positive sentiment tendency, first-order differenced neutral sentiment tendency, and first-order differenced negative sentiment tendency with the first-order differenced Shanghai Composite Index (CIPr) and trading volume (Tradvol). The corresponding equations are as follows:

$$\begin{cases} DCIPr_t = a_0 + \sum_{i=1}^{k} a_i DCIPr_{t-i} + \sum_{j=1}^{k} b_j DPositive_{t-j} + \varepsilon_t \\ DPositive_t = c_0 + \sum_{i=1}^{k} c_i DPositive_{t-i} + \sum_{j=1}^{k} d_j DCIPr_{t-j} + \varepsilon_t \end{cases} \quad (2)$$

$$\begin{cases} DCIPr_t = a_0 + \sum_{i=1}^{k} a_i DCIPr_{t-i} + \sum_{j=1}^{k} b_j DPositive_{t-j} + \varepsilon_t \\ DPositive_t = c_0 + \sum_{i=1}^{k} c_i DPositive_{t-i} + \sum_{j=1}^{k} d_j DCIPr_{t-j} + \varepsilon_t \end{cases} \quad (3)$$

$$\begin{cases} DCIPr_t = a_0 + \sum_{i=1}^{k} a_i DCIPr_{t-i} + \sum_{j=1}^{k} b_j DNeutral_{t-j} + \varepsilon_t \\ DNeutral_t = c_0 + \sum_{i=1}^{k} c_i DNeutral_{t-i} + \sum_{j=1}^{k} d_j DCIPr_{t-j} + \varepsilon_t \end{cases} \quad (4)$$

$$\begin{cases} DTradvol_t = a_0 + \sum_{i=1}^{k} a_i DTradvol_{t-i} + \sum_{j=1}^{k} b_j DNeutral_{t-j} + \varepsilon_t \\ DNeutral_t = c_0 + \sum_{i=1}^{k} c_i DNeutral_{t-i} + \sum_{j=1}^{k} d_j DTradvol_{t-j} + \varepsilon_t \end{cases} \quad (5)$$

$$\begin{cases} DCIPr_t = a_0 + \sum_{i=1}^{k} a_i DCIPr_{t-i} + \sum_{j=1}^{k} b_j DNegative_{t-j} + \varepsilon_t \\ DNegative_t = c_0 + \sum_{i=1}^{k} c_i DNegative_{t-i} + \sum_{j=1}^{k} d_j DCIPr_{t-j} + \varepsilon_t \end{cases} \quad (6)$$

$$\begin{cases} DTradvol_t = a_0 + \sum_{i=1}^{k} a_i DTradvol_{t-i} + \sum_{j=1}^{k} b_j DNegative_{t-j} + \varepsilon_t \\ DNegative_t = c_0 + \sum_{i=1}^{k} c_i DNegative_{t-i} + \sum_{j=1}^{k} d_j DTradvol_{t-j} + \varepsilon_t \end{cases} \quad (7)$$

The Granger causality test results, presented in Table 3, provide insights into the relationships between the various time series. It is observed that the variation in the positive sentiment tendency time series (DPositive) exhibits mutual Granger causality with the lagged 1-day variation in the trading volume of the Shanghai Composite Index (DTradvol). However, it does not have a significant effect on the variation in the closing price of the Shanghai Composite Index (DCIPr). This implies that the preceding positive sentiment tendency of investors influences subsequent changes in stock market trading volume, but it does not directly impact changes in stock market closing prices. On the other hand, the variation in the negative sentiment tendency time series (DNegative) is identified as the Granger cause of the lagged 1 to 5-day variation in the closing price of the Shanghai Composite Index (DCIPr). However, it does not have a significant effect on the variation in the trading volume of the Shanghai Composite Index (DTradvol). This indicates that the preceding negative sentiment tendency of investors influences subsequent changes in stock market closing prices, but it does not directly affect changes in stock market trading volume. Furthermore, the variation in the neutral sentiment tendency time series (DNeutral) is identified as the Granger cause of the lagged 1-day variation in the trading volume of the Shanghai Composite Index (DTradvol). However, it does not have a significant effect on the variation in the closing price of the Shanghai Composite Index (DCIPr). This suggests that the preceding neutral sentiment of investors does not directly influence subsequent changes in stock market closing prices, but it does impact changes in stock market trading volume. In summary, the analysis indicates that positive sentiment tendency influences trading volume, negative sentiment tendency affects closing prices, and neutral sentiment tendency impacts trading volume.

Table 3: Granger causality test on the relationship between investor sentiment tendency and stock market performance. (Notes: * denote significance for 5%)

	LD	DP	DN	DN	DC	DT
DC	1	0.526	0.698	0.036*	-	-
DC	2	0.572	0.901	0.027*	-	-
DC	3	0.582	0.937	0.007*	-	-
DC	4	0.760	0.990	0.004*	-	-
DC	5	0.832	0.988	0.002*	-	-
DT	1	0.035*	0.037*	0.716	-	-
DT	2	0.151	0.176	0.554	-	-
DT	3	0.084	0.120	0.612	-	-
DT	4	0.118	0.221	0.196	-	-
DT	5	0.131	0.148	0.096	-	-
DP	1	-	-	-	0.381	0.045
DP	2	-	-	-	0.332	0.118
DP	3	-	-	-	0.275	0.316
DP	4	-	-	-	0.255	0.106

DP	5	-	-	-	0.355	0.185
DN	1	-	-	-	0.466	0.072
DN	2	-	-	-	0.320	0.162
DN	3	-	-	-	0.343	0.396
DN	4	-	-	-	0.383	0.128
DN	5	-	-	-	0.442	0.211
DN	1	-	-	-	0.309	0.366
DN	2	-	-	-	0.888	0.577
DN	3	-	-	-	0.961	0.743
DN	4	-	-	-	0.990	0.886
DN	5	-	-	-	0.988	0.935

The original sequences of the negative sentiment tendency of investors and the closing price of the Shanghai Composite Index exhibit stationarity before first differencing. Consequently, the results of the Granger causality test for the original sequences align with the results obtained after first differencing. Therefore, it is possible to forecast the closing points of the Shanghai Composite Index based on the proportion data of the negative sentiment tendency of investors derived from stock forum posts. In contrast, the positive sentiment tendency and neutral sentiment tendency of investors do not demonstrate a consistent predictive effect on the closing prices of the stock market. This observation may be attributed to the social dynamics of stock forums, wherein users tend to be more responsive to content with a negative bias, thereby amplifying the impact of panic sentiment on the stock market.

From an economic perspective, the stock market typically operates on a weekly cycle consisting of approximately five trading days. It has been observed that changes in the negative sentiment tendency sequence of investors, occurring 1 to 5 days prior, can effectively predict changes in the closing price of the Shanghai Composite Index. Notably, the test results from 5 days prior demonstrate the highest level of significance. This finding can be interpreted as investors requiring a certain amount of time to process and assimilate negative information. After a week of careful observation and analysis, they are better able to incorporate the previous negative information and form a more conclusive judgment regarding stock prices. Consequently, they are more likely to make investment decisions aligned with their assessment on that particular day. On the other hand, the influence of changes in positive sentiment tendency and neutral sentiment on stock market performance is primarily reflected in changes in trading volume. Furthermore, their effects exhibit a shorter cycle and yield rapid outcomes. Specifically, changes in positive sentiment tendency and neutral sentiment from the previous day directly impact changes in trading volume.

4 CONCLUSIONS

This research paper focuses on the period from January 1, 2022, to April 28, 2023, and aims to examine the impact of investor sentiment on the performance of the Shanghai Composite Index. Specifically, the study investigates the influence of three types of sentiment (positive, neutral, and negative) extracted from the analysis of stock forum posts on the closing price, trading

volume, and returns of the index. The findings reveal that different sentiment tendencies have distinct effects on the stock market. The analysis demonstrates that changes in positive and neutral sentiment, as identified from the analysis of stock forum posts, have a significant impact on the trading volume of the Shanghai Composite Index. These sentiments evoke strong but short-lived responses from investors, leading to immediate considerations of buying or selling stocks within a brief timeframe. In contrast, changes in negative sentiment exert a notable influence on the closing price and returns of the index. Negative sentiment has a longer-lasting stimulating effect on investors, requiring them to gradually process and react to negative information. This finding aligns with the reality of the stock market, where investors' investment decisions are primarily driven by negative sentiment. It implies that investors need time to assimilate and appropriately respond to negative information. These results underscore the significance of rational endorsement and critical considerations in the stock market. The market's existence and long-term development rely on investors' acceptance and recognition, which are rooted in thoughtful evaluation rather than blind enthusiasm.

Individual investors have a significant presence in China, making the stock market more susceptible to the influence of investor sentiment. Therefore, effectively mining investor sentiment is a key focus in studying the relationship between investor sentiment and stock market performance. This paper can serve as a valuable supplement to validate the rationality of behavioral finance theories in the era of big data. Based on the findings of this study, a predictive model for the closing price of the Shanghai Composite Index can be established to provide investment references for investors. Additionally, the research in this paper demonstrates that stock forums can serve as a platform for guiding investor sentiment. Managers can utilize this platform for sentiment management and early warning, enabling them to take proactive measures to address potential financial market panics or riots. This can effectively guide rational investments by stockholders and establish a well-regulated and orderly stock market environment.

REFERENCES

[1]Black F. Noise [J]. Journal of Finance, 1986(41):529-543.

[2]Delong J B, Shleifer A, Summers L H, et al. Noise trader risk in financial markets [J]. Journal of Political Economy, 1990, 98(4): 703-738.

[3]Ba Shusong, Zhu Hong. Margin financing, investor sentiment and market volatility [J]. Studies of International Finance, 2016(8): 82-96.

[4]Wen Fenghua, Xiao Jinli, et al. The effects of characteristics of investor sentiment on stock price behaviours [J]. Journal of Management Sciences in China, 2014, 17(3): 60-69.

[5]Bu H, Pi L. Does investor sentiment predict stock returns? The evidence from Chinese stock market [J]. Journal of Systems Sciences and Complexity, 2014, 27(1): 130-143.

[6]Baker M, Wurgler J. Investor sentiment and the cross-section of stock returns [J]. Journal of Finance, 2006, 61(4): 1645-1680.