An Explanatory Model for the Impact of Investor Sentiment on Stock Markets Integrating XGBoost and SHAP

Yuhao Wang*, #, Yifei Yun#, Qijie Wang#

School of Management, Xiamen University, Xiamen, China, 361005

* Corresponding Author Email: 17420212201029@stu.xmu.edu.cn

#These authors contributed equally.

Abstract. With the widespread popularity of Internet, non-professional individual investors can share information and express their tendency through online forum, exerting a certain influence on the price fluctuations in the stock market. This study has compiled investor sentiment data from the stock forum of Eastmoney.com for the years 2018 to 2022. By integrating this data with the daily return rates of individual stocks in the Shanghai Stock Exchange (SSE), an XGBoost regression model has been established. Utilizing SHAP (Shapley Additive exPlanations) for visualization and analysis, the findings reveal that the daily stock returns exhibit a significant positive correlation with investor positive sentiment and turnover rate, and a significant negative correlation with the number of forum posts. Conversely, there is no significant correlation with metrics such as average number of followers, average reading volume, average number of likes, average net comments, user average influence index, and emotional consistency index within the stock forum. Moreover, in the Chinese SSE market, stocks of different market capitalizations do not display significant differences in sensitivity to investor sentiment. This study contributes novel perspectives and methodologies to the study of investor sentiment, aiding in a more profound comprehension of the behaviors of individual investors participating in stock forum discussions and the consequential impact of such behaviors on the market.

Keywords: XGBoost, SHAP, Investor Sentiment, Stock return, Online Forum.

1. Introduction

The Chinese stock market has long been regarded as immature due to the disproportionately high participation of individual investors. In the A-share market, individual investors contribute to 80% of trading volume. However, influenced by subjective factors such as life experiences, social backgrounds, knowledge, and personality preferences, individual investors' behaviors in investment activities do not strictly adhere to rational assumptions, reflecting a state of incomplete rationality. When pricing assets based on their expectations of future returns and risks, individual investors, impacted by psychological factors like overconfidence and herding behavior, may exhibit cognitive biases that lead to distorted expectations, termed as investor sentiment.

Moreover, due to incomplete market trading mechanisms in China, such as restrictions on short selling and the immaturity of institutional investors, arbitrage opportunities may not be promptly realized. Hence, the asset pricing issue in the Chinese stock market merits thorough exploration, especially regarding how investor sentiment influences asset pricing.

To investigate these issues, many domestic scholars have adopted various sentiment measurement methods and constructed different sentiment indicators. However, there is no consensus on the impact of investor sentiment on the market. Some researchers have found a significant positive correlation between investor sentiment and same-day returns [1], while other empirical studies suggest a significant negative correlation or no correlation [2]. Therefore, further research on the impact of investor sentiment on the Chinese stock market is warranted.

To address these research gaps, this study collects investor sentiment data from the Eastmoney.com stock forum (2018-2022). Integrating this data with the daily return rates of individual stock in the SSE market, this study constructs an XGBoost regression model. Utilizing

SHAP for visualization and analysis, this study explains the significance and positive/negative impact of various factors on stock returns.

The innovations of this study lie in the following aspects: in terms of research perspective, besides selecting the number of forum posts, investor positive sentiment index, and emotional consistency index as in previous studies, this research also considers various feature data reflecting investor sentiment, such as average number of followers and average reading volume, expanding the scope of domestic research; in terms of research methods, this study adopts the XGBoost model and the SHAP interpretable method based on machine learning, validating and complementing existing research conclusions and providing ideas and methods for investor sentiment research; in terms of research content, this study groups stocks based on company market capitalization, exploring the influence of different investor sentiment features on the intraday return rate of individual stocks, furnishing regulatory authorities with insights for informed decision-making in market control and supervision.

2. Literature Review

2.1. Impact of Investor Sentiment on Stock Markets

Behavioral finance posits that investor sentiment is a crucial factor influencing stock prices. Barberis et al. [3] argue that two psychological biases affect investor decisions. Firstly, due to a conservative inclination, investors may fail to promptly adjust their views on stock prices based on new information, resulting in inadequate market responses to such information and a positive impact of investor sentiment on future stock prices. Secondly, the representativeness heuristic bias may lead to an overestimation of the probability of random events, causing an excessive market reaction to new information and a negative impact of investor sentiment on future stock prices. Baker and Wurgler [4] contend that different types of stocks display varying sensitivity to investor sentiment. Smaller market capitalization, shorter establishment time, higher stock volatility, or loss-making listed companies are more susceptible to investor sentiment, whereas large-cap stocks with stable profit levels exhibit characteristics akin to bonds and are less influenced by investor sentiment.

Domestic scholars have also delved into this area. Duan Jiangjiao found a positive correlation between investor sentiment and same-day stock returns but no significant impact on returns over the following two days. Moreover, they discovered a negative correlation between same-day forum post volume and stock returns over the subsequent two days. Chen Rongda explored investor sentiment in internet finance, revealing a negative correlation between this indicator and returns on internet financial products. Qian Yu [5] studied the impact of sentiment consistency among forum users on returns and volatility. Yao Jiaquan [6] constructed a financial sentiment dictionary and found that these sentiment indicators based on annual reports and social media texts significantly predicted stock returns, volatility, and the risk of stock price collapse. Fan Xiaoyun [7] used an N-gram model to build a sentiment dictionary, discovering that stock forum sentiment significantly predicted stock market returns, trading volume, and volatility, with minimal impact on macroeconomic factors. Zhang Ke [8] established sentiment indicators for opinion leaders and non-opinion leaders, revealing that opinion leaders have some predictive power for stock returns, as investors are more prone to overlook bearish opinions from opinion leaders. Li Helong [9], through the analysis of 115 core studies, identified asymmetric effects of social media sentiment with different orientations on returns. Ni Wenhui [10] demonstrated that investor consensus has a significantly positive impact on stock returns, and this impact persists due to the emergence of crowded buying behavior stimulated by investor consensus, leading to asset price increases.

From the above literature, it is evident that both domestic and international scholars have conducted substantial research on investor sentiment. However, the predominant approach has been to measure investor sentiment as a collective entity, and the chosen indicators for constructing sentiment measures have not been sufficiently comprehensive. Critical features such as investor influence and average post reading volume have been overlooked. Furthermore, there is still no consensus on the impact of investor sentiment on the market, and a comprehensive study on the

sensitivity of stocks with different market capitalizations to investor sentiment in the Chinese market remains lacking. Therefore, further research in this area is warranted.

2.2. Application of the SHAP Model in Information Resource Management

The SHAP explanation method, proposed by Lundberg et al. [11], is a form of attribution-based interpretation designed to elucidate black-box models. SHAP values are employed as the attribution values for features, primarily quantifying the contribution of each feature to model predictions. Compared to traditional feature importance methods, the SHAP explanation method demonstrates better consistency, revealing the positive or negative relationships of each feature factor with the target variable. It can be applied for both local and global interpretation. In local interpretation, each feature has its set of SHAP values, allowing for the explanation of each feature's contribution to predictions for every sample. For overall interpretation, the average of the SHAP values for that feature across all samples is computed as the feature's importance value. The greatest advantage of the SHAP model lies in its ability to reflect the influence of each feature within each sample and the positivity or negativity of this impact on the final prediction result.

The SHAP model has gradually found application in various fields such as network sentiment analysis and social network user behavior. Jiang Jianhong et al. [12], for instance, utilized SHAP to visualize the impact mode and magnitude of machine learning on cognitive evaluation metrics under negative events for Weibo users, providing relevant recommendations. Sun Ran et al. [13] employed the SHAP interpretation method to rank feature importance in studying linguistic behavior of social media users during emergency events. Li Weiqing et al. [14] conducted interpretable analysis using the SHAP method, exploring the influence of perceived value dimensions on the continuous usage behavior of users on a barrage video website.

While the SHAP model has demonstrated promising results in the above-mentioned domains for assessing the contribution of predictive results, its application in investigating factors influencing stock returns in the stock market remains relatively limited. Considering the central issue of this study, "The Impact of Investor Sentiment on Stock Returns," the introduction of the SHAP interpretable machine learning model in this domain could enhance model interpretability, offering new perspectives and methodologies for the study of investor sentiment.

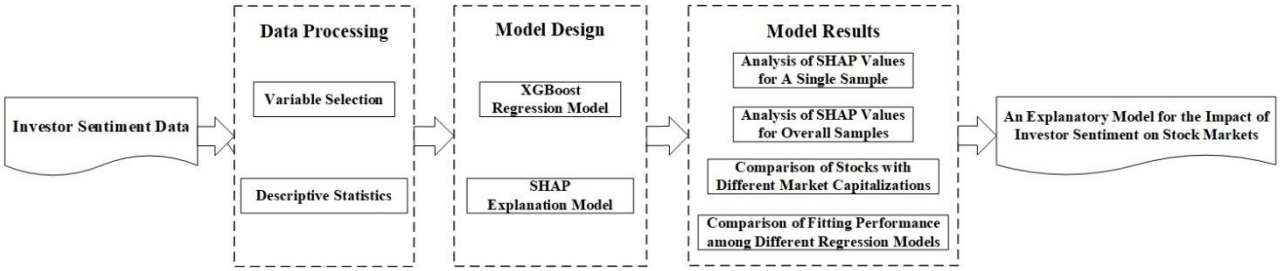


Figure 1. An Explanatory Model for the Impact of Investor Sentiment on Stock Markets Integrating XGBoost and SHAP

3. Data Preprocessing

3.1. Data Source

The specific research framework of this study is illustrated in Figure 1. The dataset utilized in this research comprises stock data from the Chinese SSE market retrieved from the CSMAR database, spanning from January 1, 2018, to December 31, 2022. After excluding data related to suspended trading and financial sector stocks, a total of 506,660 records were retained, encompassing 772 A-share companies on the SSE market.

3.2. Variable Selection and Descriptive Statistics

Table.1 presents the chosen dependent and independent variables, along with their respective definitions or formulas. Table.2 provides the descriptive statistics for these variables.

Table 1. Variable Names and Their Definitions or Formulas

Variable	Description
Change Ratio	$\frac{(\text{Close-Open})_{\text{today}}}{\text{Prev close}} \times 100\%$
X_1	The total number of posts in the company's forum during a specified time period.
X_2	The average number of followers of the poster.
X_3	The average number of readings/views per post
X_4	The average number of comments on posts excluding the poster's own comments.
X_5	The average number of thumbs-ups received by all posts.
X_6	Average influence index of users posting in the forum.
X_7	$\frac{(\text{BullishPosts-BearishPosts})}{(\text{BullishPosts+BearishPosts})}$
X_8	$1-\sqrt{(1-X_7^2)}$
X_9	The turnover ratio of a stock during a specified time period.

Table 2. Descriptive Statistics for Variables

Variable	Maximum	Minimum	Mean	Median	Variance
Change Ratio	50	-12.5	0.085321	0	7.23
X_1	10528	0	30.44773	13	5217.67
X_2	923012	0	9272.621	15.38	4675202647
X_3	3403749	0	3177.452	403.095	2107148756
X_4	3519	0	5.697803	0	6690.79
X_5	4945	0	8.360196	1.19	15514.98
X_6	5	0	2.148747	2.15	0.63
X_7	1	-1	-0.1289	-0.14286	0.19
X_8	1	0	0.157141	0.037909	0.09
X_9	743.86717	0.00079	16.90908	9.186274	591.48

The aforementioned table reveals a consistently negative sentiment in the Eastmoney.com stock forum, as indicated by both the mean and median of the sentiment index. This underscores a systematic inclination among contributors to express pessimistic views. Furthermore, it suggests that individual investors tend to harbor negative emotions, preferring to articulate them within the online platform. Additionally, the substantial variance in variables such as average number of followers, average reading volume, average number of likes, average net comments emphasizes the significant diversity in emotional expressions among investors. It implies the challenge of adequately characterizing investor sentiment using a single indicator. Based upon previous research, this study integrates these diverse indicators to comprehensively portray investor sentiment.

4. Model Design

4.1. Introduction to Basic Concepts of XGBoost

XGBoost is a machine learning algorithm based on decision trees widely applied in the field of information resources for regression and classification tasks. In this algorithm, decision trees are sequentially associated, with each subsequent prediction building upon the errors of the preceding

round. Through iterative refinement utilizing prediction errors from each round, the model is constructed to enhance predictive accuracy. XGBoost is known for its efficient scalability and robustness, making it a preferred choice in various applications.

4.2. Construction of the XGBoost Regression Model

$$ChangeRatio = \beta_0 + \sum_{i=1}^{9} \beta_i X_i + \varepsilon \quad (1)$$

The regression model, as denoted in Equation (1), is established in this study, where $\beta_i (i = 1, 2...9)$ represents the correlation coefficient, and ε signifies the errors term. Various independent variables encompass dimensions associated with both posts and user behaviors, offering a comprehensive examination of the impact of stock forum sentiment factors on stock price changes.

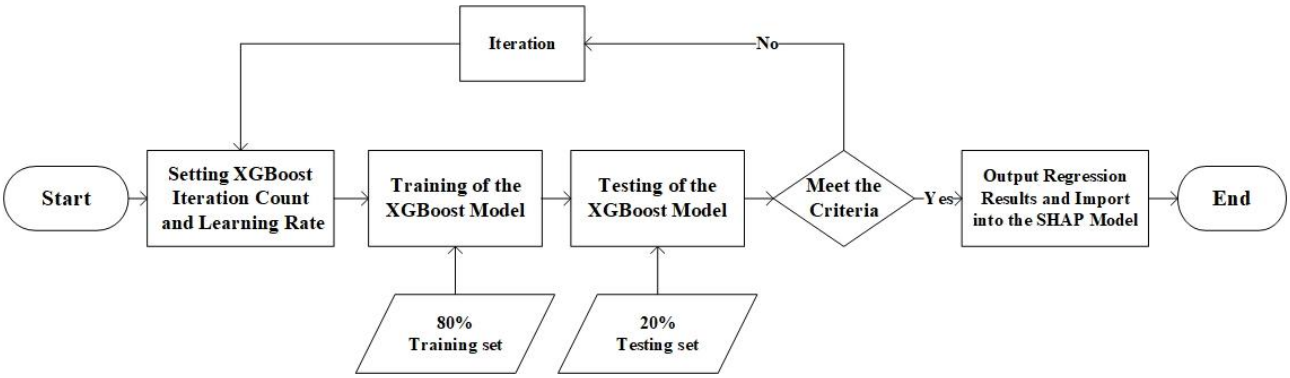


Figure 2. Workflow for Constructing the XGBoost Regression Model

As illustrated in Figure 2, the subsequent steps involve utilizing 80% of the data as the training set and 20% as the testing set. The XGBoost learning rate is updated, and the iteration count is set. The investor sentiment and stock return sample training data are then input into the model for fitting. Generalization ability is validated on the sample testing set to assess whether the model meets the required criteria. If the criteria are not met, an iterative approach is employed to search for improved parameter values. Ultimately, the optimized regression fitting results are output, and the SHAP model's interpreter function is applied. Lastly, utilizing the SHAP interpretable method based on machine learning, we explore the impact levels of these independent variables.

5. Model Results

5.1. Analysis of SHAP Values for A Single Sample

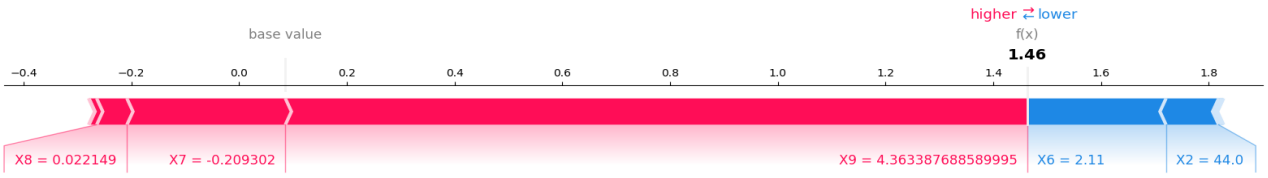


Figure 3. Contribution of Individual Features in the XGBoost Model for a Single Sample

Utilizing the SHAP method, the machine learning model, trained beforehand, is employed to calculate the feature contributions for each sample, resulting in the corresponding SHAP values. The sum of all feature SHAP values for a particular sample represents the model's prediction for that sample. Taking the stock of LLIAONING SHIDAI WANHENG CO.,LTD (Code: 600241) as an example, an XGBoost model is employed, and under the SHAP method, the visualization of the contribution of each feature to the stock returns is illustrated in Figure 3. In this representation, blue denotes features with a negative impact on stock returns, while red indicates features with a positive impact. For this stock, apart from the negative impact of average number of followers and user

average influence index, all other features positively influence its returns. Among them, turnover rate exhibits the most significant positive impact on stock returns, followed by the investor sentiment index. Furthermore, considering the comprehensive impact of all features on this sample, the predicted return value is 1.46%, signifying an overall positive influence of all features on returns, and this closely aligns with the average actual stock return of 1.42% over the period from 2018 to 2022. To a certain extent, this confirms the correctness of the selected features in this study and the effectiveness of the SHAP method.

5.2. Analysis of SHAP Values for Overall Samples

5.2.1 Analysis of Feature Importance

Applying the SHAP model, the SHAP values of various features for the overall sample are obtained, as shown in Figure 4. It is observed that investor sentiment index, turnover rate, and the number of forum posts contribute significantly to stock returns, while average number of followers, average reading volume, average number of likes, average net comments, user average influence index, and emotional consistency index contribute less and exhibit lower significance.

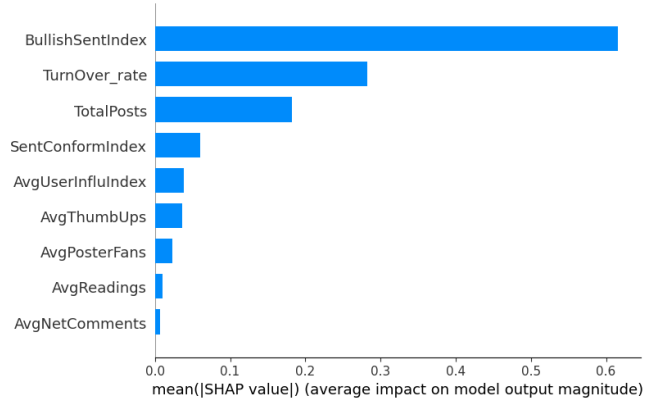


Figure 4. The SHAP Values of Various Features for the Overall Samples

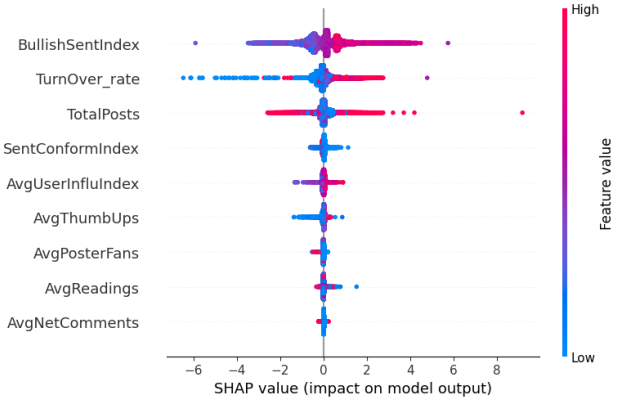


Figure 5. A Scatter Plot of SHAP Values

5.2.2 Analysis of Feature Positivity/Negativity

Visualizing the values of each feature across all samples and constructing a scatter plot of SHAP values, as depicted in Figure 5, this study further explores the positivity or negativity of each feature's impact on viewership. In this representation, red indicates high SHAP values for the corresponding feature on a sample, while blue indicates low SHAP values. If a feature is positively correlated with stock returns, its representation in the scatter plot exhibits a transition from blue to red, with the zero SHAP value as the dividing line (blue on the left, red on the right, and purple near the zero SHAP value); conversely, if it is negatively correlated, the representation transitions from red to blue.

As illustrated in Figure 5, investor sentiment index and turnover rate are positively correlated with stock returns, while the number of forum posts is negatively correlated. Due to the smaller contribution of average number of followers, average reading volume, average number of likes, average net comments, user average influence index, and emotional consistency index, the subsequent analysis primarily focuses on the top three indicators and their mechanisms influencing stock returns.

Based on the theories of investor sentiment, combined with relevant literature on stock forum discussions, we posit that when the number of forum posts for a stock increases, it gains more attention, resulting in a rise in bullish sentiment and an increased likelihood of being purchased by individual investors. Similarly, an increase in turnover rate for a stock signifies amplified trading volume, suggesting that investors are buying in significant quantities. In the absence of short-selling restrictions, this leads to a short-term increase in stock prices, resulting in positive stock returns. On the other hand, when the number of forum posts increases, it serves as intuitive information, and investors tend to overreact to easily processed information. Due to speculative characteristics such as chasing trends, short-term trading, and frequent turnovers prevalent among Chinese investors, stocks

highly discussed in forums easily become excessively focused upon. Rational investors may recognize the presence of attention premiums in stock prices, leading to an increase in bearish sentiment in forum posts. Rational investors may sell stocks that are excessively focused on. As post sentiment turns bearish, attention shifts for individual investors, with no further buyers following, and stocks with a higher number of forum posts experience larger price declines. Thus, short-term stock returns rise with increasing post sentiment and fall with decreasing post sentiment.

5.3. Comparison of Stocks with Different Market Capitalizations

Baker and Wurgler [2] posited that stocks with different market capitalizations exhibit varying sensitivities to investor sentiment. This study utilizes the SHAP model to investigate whether such differences exist. The 722 listed A-share companies on the Chinese SSE market are divided into two groups based on the average market capitalization (23.26 billion), and the SHAP values and scatter plots for various features are presented in Figures 6–8.

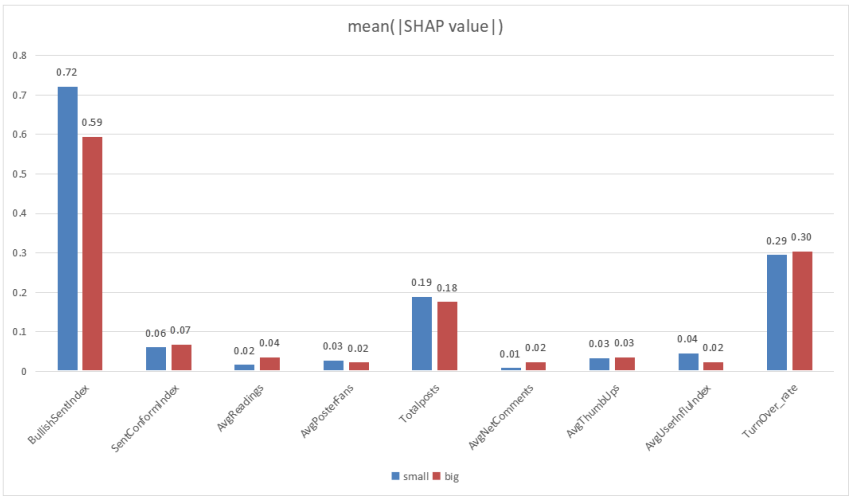


Figure 6. The SHAP Values for Stocks with Different Market Capitalizations

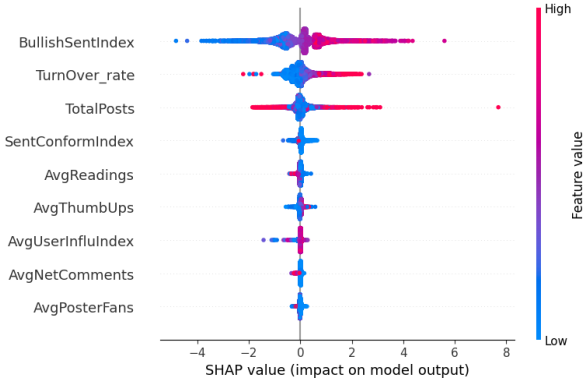


Figure 7. A Scatter Plot of SHAP values for Stocks with Larger Market Capitalization

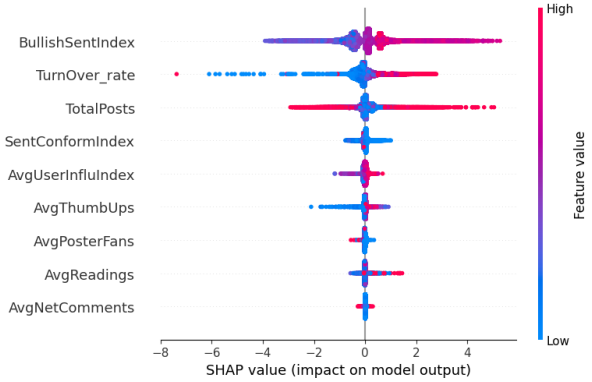


Figure 8. A Scatter Plot of SHAP values for Stocks with Smaller Market Capitalization

As observed in Figure 7, the SHAP values for factors affecting stock returns are nearly identical for large and small market capitalization companies. Therefore, it can be inferred that there is no significant difference in the sensitivity of stocks with different market capitalizations to investor sentiment. Possible reasons include the different stock market conditions and indicators used to characterize investor sentiment in this study compared to previous research. The Chinese SSE market is relatively complex, with various factors influencing stock returns. Figures 7 and 8 demonstrate that, regardless of whether stocks have large or small market capitalizations, investor sentiment index and turnover rate are positively correlated with stock returns, while the number of forum posts is negatively correlated, further validating the conclusions drawn in Section 5.2.2.

5.4. Comparison of Fitting Performance among Different Regression Models

In order to assess the fitting performance of the XGBoost regression model, this study employs three evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2). The formulas for these three evaluation metrics (Equations 2-4) are defined as follows:

$$MAE = \frac{1}{N}\sum_{n=1}^{N}| y_n - \hat{y}_n | \tag{2}$$

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 \tag{3}$$

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{\hat{y}}_n)^2} \tag{4}$$

Here, y_n, $\hat{y}_n$ and $\bar{y}_n$ represent the actual values, predicted values, and mean values of stock returns, respectively. MAE and MSE measure the errors between predicted and actual values, with lower values indicating better performance. R^2 describes the model's ability to fit the data, and a value closer to 1 indicates better fitting.

Table 3. Comparison of Fitting Performance among Different Regression Models

Regression Models	Training Set(80%)			Testing Set(20%)		
	MAE	MSE	R^2	MAE	MSE	R^2
Multiple Linear Regression	1.796	6.654	0.081	1.792	6.646	0.078
KNN	**1.600**	**5.325**	**0.265**	1.953	7.970	-0.106
AdaBoost	1.784	6.429	0.112	1.781	6.454	0.105
XGBoost	1.645	5.435	0.249	**1.676**	**5.719**	**0.207**

Utilizing 80% of the data for training and 20% for testing, the fitting performance of the four regression models is presented in Table.3. The bold values represent the optimal performance, demonstrating that XGBoost exhibits the best fitting performance on the testing set.

5.5. Analysis of Experimental Conclusions

With respect to the investor sentiment data from the Eastmoney.com stock forum and daily stock return data, this study extracted nine independent variables to comprehensively characterize investor sentiment. Descriptive statistics reveal significant variability in most variables, indicating substantial differences in investor sentiment among different investors, making it challenging to capture investor sentiment with a single indicator. Subsequently, the dataset was divided into training and testing sets, and an XGBoost regression model was established. Finally, utilizing the SHAP model, the study investigated the influence of investor sentiment on stock returns from dimensions such as individual sample feature values, overall sample feature values, and stocks with different market capitalizations, enriching the application of this model in stock market research. Comparisons of different model regression fitting performances show that the XGBoost model exhibited the best performance in terms of MAE, MSE, and R^2 on the testing set, enhancing the robustness of the experiment.

6. Conclusion

6.1. Research findings

This study utilized an XGBoost-based SHAP interpretable model to investigate the impact of investor sentiment on stock returns. The main research findings are as follows:

(1) The same-day stock return is significantly positively correlated with daily investor positive sentiment and turnover rate while being significantly negatively correlated with the number of forum posts. This implies that daily investor positive sentiment, turnover rate, and the number of forum posts have predictive power for daily stock returns.

(2) There is no significant correlation between same-day stock returns and average number of followers, average reading volume, average number of likes, average net comments, user average influence index, and emotional consistency index within the stock forum. This indicates that these variables do not have good predictive power for stock returns.

(3) In the Chinese SSE market, there is no difference in the sensitivity of stocks with different market capitalizations to investor sentiment.

6.2. Research Suggestions and Discussion

This study offers several insights: in the absence of new significant news or fundamental information about a company, the influence of stock forums on the stock market cannot be ignored. As a gathering place for a large number of investor opinions, stock forums provide a direct and effective way to measure investor sentiment and attention. Regulatory authorities in the securities market can use this information to enhance regulatory efficiency: monitoring forum posting volume. Securities market regulatory authorities can obtain advance signals of future stock price fluctuations in the relatively easy-to-monitor environment of stock forums and take appropriate measures before significant stock price fluctuations occur to prevent the risk of sharp rises and falls in stock prices; monitoring forum post sentiment. From the impact of post sentiment on returns, when investor sentiment and turnover rate increase, it indicates that irrational investor sentiment is prevalent. Through sentiment monitoring, the risk of stocks being manipulated in forums can be prevented.

Limitations of this study include the selection of A-share companies on the Chinese SSE market as the research objects. Future research can selectively choose stocks from the Shenzhen market to expand the data sample. Additionally, various channels such as news text, social media, and investor interaction platforms can be considered to comprehensively characterize investor sentiment, providing a more comprehensive reflection of its impact on stock returns.

References

[1] Duan Jiangjiao, Liu Hongzhong, Zeng Jianping. Analysis on the information content of China's internet stock message boards[J]. Journal of Financial Research, 2017, (10) : 178 -192.

[2] Chen Rongda, Lin Bo, He Chengying, et al. Internet finance characteristics, internet finance investor sentiment and the return of internet financial products[J]. Economic Research Journal, 2019, 54(7) : 78-93.

[3] Barberis N, Shleifer A, Vishny R. A model of investor sentiment[J]. Journal of Financial Economics, 1998,49(3):307-343.

[4] Baker M, Wurgler J. Investor sentiment in the stock market [J]. Journal of Economic Perspectives, 2007,21(2): 129 -151.

[5] Qian Yu, Li Zirao, Li Qiang, et al. Impact of online community support tendencies on returns and volatility in Chinese stock market[J]. Journal of Management Sciences in China, 2020, 23(2) : 141-155.

[6] Yao Jiaquan, Feng Xu, Wang Zanjun, et al. Tone, sentiment and market impacts: The construction of Chinese sentiment dictionary in finance[J]. Journal of Management Sciences in China, 2021, 24(5) : 26 -46.

[7] Fan Xiaoyun, Wang Yedong, Wang Daoping, et al. Heterogeneity analysis of information content for financial text from different sources: A hybrid text sentiment measurement method[J]. Journal of Management World, 2022, 38(10):78-101.

[8] Zhang ke, Li Xindan, Fang Xiao, et al. Opinion leader at the online stock forum: Having the discerning eve or eve-catching[J]. Journal of Management Sciences in China.2022, 25(9):90-107.

[9] Li Helong, Ren Changsong, Liu Xinru, et al. A Review of Textual Sentiment Research in Financial Markets[J]. Data Analysis and Knowledge Discovery, 1-25[2023-3-27].

[10] Ni Wenhui, Zhang Pu. The Influence of Investors' Opinion Convergence on Stock Market ——Based on the Dual Perspective of Return and Risk [J]. Journal of Financial Development Research, 2023(04):18-28.

[11] Lundberg S, Lee S I. A Unified Approach to Interpreting Model Predictions[C] / / Conference and Workshop on Neural Information Processing Systems. California: NIPS Press, 2017:4765-4774.

[12] Jiang Jianhong, Li Mengxin. An Explanatory Model for Weibo Users' Extreme Emotions under Negative Events of Enterprises Integrating LGBM and SHAP[J]. Data Analysis and Knowledge Discovery, 1-24[2023-10-18].

[13] Sun Ran, An Lu. The Speech Acts Classification of Social Media Users in the Context of Public Emergencies[J]. Journal of Information Resources Management 1-12[2023-10-24] .

[14] Li Weiqing, Wang Wentao, Huano W, et al. Study on the Influencing Factors and Its Explainability of Users' Continuous Use Behavior of Bullet Screen Video Website—Perspective Based on Perceived Value[J]. Journal of Modern Information, 1-10[2023-06-21].