The Impact of Investor Sentiment on Stock Returns

Ao Huang[1] and Ruiqi Huang [2,*]

[1] School of Mathematical Sciences, Tongji University, 200092 Shanghai, China
[2] School of Mathematics, Xi'an Jiaotong-Liverpool University, 215000 Suzhou, China
*Corresponding author: ruiqi.huang22@student.xjtlu.edu

Abstract. Currently, the application of alternative media data offers a new perspective for measuring investor sentiment. To better predict stock returns and achieve excess returns, this paper reviews both theoretical research on the impact of investor sentiment on stock returns and algorithms for measuring investor sentiment through alternative media data. Firstly, this paper provides a review of research conducted by scholars from various countries and eras on the effect of various types of sentiment on stock performance. It then introduces two distinct algorithms from the perspectives of traditional news media and social media, respectively, used to extract sentiment from media text and images to gauge investor sentiment reflected in media data. Finally, empirical tests are conducted to examine the relationship between investor sentiment, as derived from different algorithms, and stock returns. The results indicate that different types of investor sentiment have a significant impact on stock returns, and incorporating investor sentiment into traditional stock return prediction models can significantly improve forecasting accuracy. By summarizing the research on alternative data and the influence of investor sentiment on stock performance, this study provides investors with a comprehensive overview of stock return prediction channels, aiding in optimizing investment strategies and achieving higher actual returns.

Keywords: Investor sentiment, Stock returns, Media data, Investment strategy, Excess returns.

1 Introduction

For decades, China's securities market has developed rapidly, but it also exhibits distinctive features compared to developed economies. According to data from the China Securities Depository and Clearing Corporation Limited (CSDCC), as shown in Figure 1, 99.77% of traders in 2023 were individual investors. From the perspective of behavioral economics, individual investors are more susceptible to emotional influences in their decision-making compared to institutional investors. In the DSSW model proposed by DeLong et al., investors who lack absolute rationality and have biased perceptions of asset prices are termed "noise traders." Unlike rational investors, inexperienced noise traders often blindly follow others' actions, lacking their own investment insights and are influenced by market conditions, rumors, and other factors to make adjustments [1]. When errors made by investors show

J. Lu (ed.), *Proceedings of the International Workshop on Navigating the Digital Business Frontier for Sustainable Financial Innovation (ICDEBA 2024)*, Advances in Economics, Business and Management Research 315,
https://doi.org/10.2991/978-94-6463-652-9_56

correlation, they are likely to make similar erroneous judgments about the same targets, thereby generating investor sentiment. Early on, the behavioral finance school introduced psychological perspectives and research methods into economics, attempting to provide rational explanations for these anomalies, which promoted the development of behavioral finance. Meanwhile, since media is not a completely neutral information intermediary, it often shows particular stances and biases when reporting news or expressing opinions, thereby conveying an implicit opinion. In this environment, individual investors are more prone to herding effects and are more likely to let news media information or the emotions of other investors on social media influence their investment decisions, which in turn affects stock market prices [2]. Thus, text data from media has become a new method for measuring investor sentiment in recent years. Given this context, this paper primarily studies the relationship between investor sentiment obtained from analyzing alternative media data and stock returns, aiming to achieve excess returns by moving beyond traditional financial data constraints in an era of increasing market transparency. The paper focuses on different algorithms for extracting investor sentiment from media data based on traditional theoretical foundations and empirically tests the impact of the obtained investor sentiment on stock returns to draw conclusions.

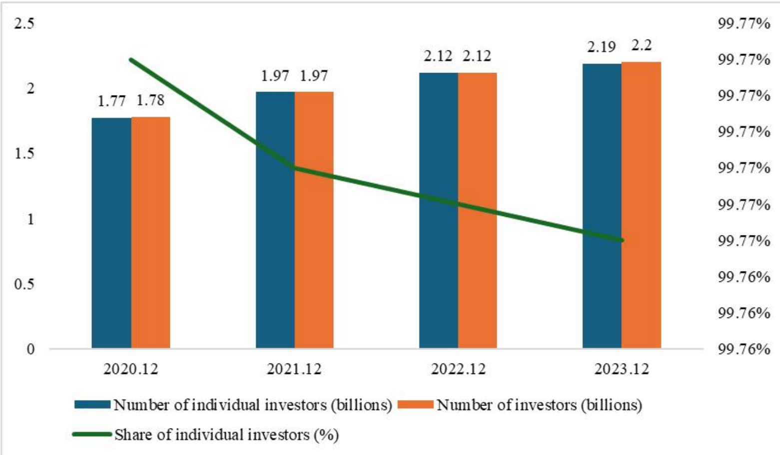

Fig. 1 Number of investors in China's securities market, December 2020-2023

2 Literature Review

In examining how investor sentiment affects stock returns, researchers have used a range of methodologies to analyze how different types of stocks respond to different sentiments, providing their own explanations for the observed variations. Malcolm Baker and Jeffrey Wurgler investigated how investor sentiment influences the differences in stock returns across various stocks. By summarizing the fluctuations in U.S. market sentiment from 1961 up to the dot-com bubble era, they concluded that when sentiment proxies are Low, stocks with small market capitalizations, newly established companies, high-volatility stocks, companies with negative earnings, non-dividend-paying equities, stocks with rapid growth and financially troubled stocks generally show relatively higher returns in the following periods. Conversely, When sentiment is elevated, these types of stocks often see comparatively lower returns in the following periods [3]. Building on the investor sentiment index proposed by Malcolm Baker and Jeffrey Wurgler, G. Mujtaba Mian et al. investigated stock price responses to earnings shocks and validated the hypothesis that pervasive market-wide investor sentiment influences how the stock market reacts to good and bad news, aligning with the direction of sentiment. They also found that the relationship between market sentiment and Stock price reactions are more pronounced for small-cap stocks, newly established companies, high-volatility stocks, non-dividend-paying equities, and distressed stocks [3, 4]. Regarding the relationship between investor sentiment and stock returns in the Chinese market, direct measurement of investor sentiment is challenging due to the confidentiality of individual investor accounts. To address this, Chi et al. used mutual fund flows as a proxy for investor sentiment across different stocks and focused on the relationship between investor sentiment and Chinese stock market returns and volatility [5]. Contrary to previous research, they found evidence suggesting that stocks with high sentiment yield higher returns compared to those with low sentiment. The researchers inferred that this might be due to the Chinese stock market being an emerging capital market where investors, having less experience, are more significantly influenced by investor sentiment [5]. Building on this research, Zhang Bo et al. explained the mechanisms behind investor sentiment in the Chinese stock market. As illustrated in Figure 2, there are three main paths: Road 1 involves Market volatility adversely impacts investor sentiment via expectations in the market.; Road 2 involves Economic cycle fluctuations and market returns enhance investor sentiment by affecting the value of market investments, with mutual impacts between market returns and the value of market investments.; Road 3 shows a direct positive relationship between market returns and investor sentiment, also indicating a reinforcing feedback loop involving market returns, market investment value, and investor sentiment. As market returns rise, the value of market investments increases correspondingly, resulting in a boost in investor sentiment, which subsequently drives further increases in market returns [6].

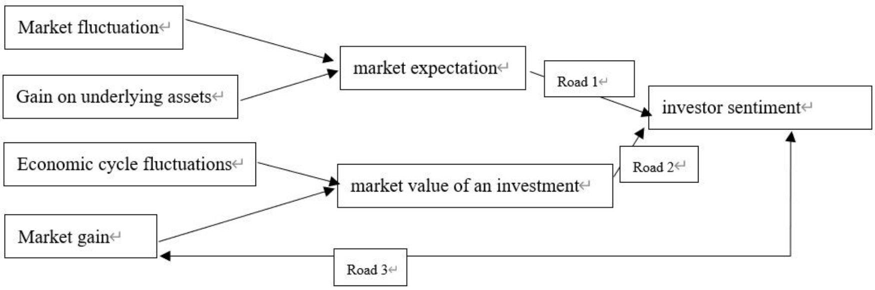

Fig. 2 Mechanisms Behind Investor Sentiment in the Stock Market

Shen Yue summarized the impact of investor sentiment on stock return fluctuations under optimistic and pessimistic scenarios, particularly focusing on the influence of external investors' behavior. As shown in Figure 3, in a period of sustained optimism, the positive investment atmosphere gradually spreads, leading to an increased acceptance of risk among market participants. Investors ramp up their investments in anticipation of higher returns. Additionally, external investors enter the market, driving up stock prices and causing stock returns to rise continuously. This sustained increase in stock returns further stimulates investor enthusiasm, creating a positive feedback loop where stock prices spiral upward, often exceeding their actual value. Conversely, during prolonged pessimism, investors become risk-averse. Panic spreads, leading to frequent withdrawals of investments, which causes stock prices to decline and stock returns to shrink. The ongoing drop in stock prices worsens investor pessimism, and this negative feedback loop is particularly pronounced in China's less mature market [7].

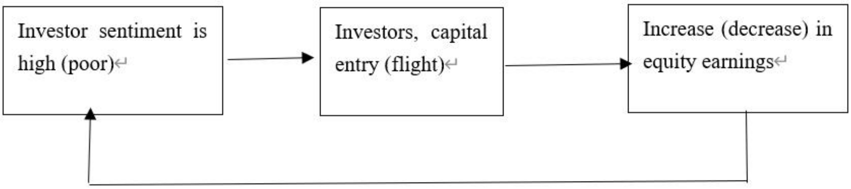

Fig. 3 The Impact of Investor Sentiment on Stock Returns

3 Existing Methods for Analyzing Investor Sentiment from Media Text Data

In the past, methods for analyzing investor sentiment were divided into direct survey methods and indirect index analysis. Recently, with the widespread study of alternative data, researching investor sentiment through alternative data reflected in media has become a new research direction. Most of the research focuses on text data, and mainstream methods involve using machine learning algorithms to analyze text data. Linguistic analysis methods, using sentiment dictionaries to measure investor sentiment in social media and news texts, have also become a new research trend. However, existing methods for constructing sentiment dictionaries still have certain limitations. Manual differentiation methods largely rely on the expertise of researchers, and the baseline word libraries used for word-matching cannot identify new internet slang or financial terminology. Additionally, sentiment dictionaries often lack coverage for emojis, which can introduce bias into sentiment measurement. Therefore, considering the different language habits of news media and social media users, it is particularly necessary to use efficient algorithms for sentiment classification and to construct targeted sentiment dictionaries.

3.1 Analyzing Investor Sentiment Through Traditional News Media

There are two main methods for measuring media sentiment. The first method is linguistic analysis, which involves breaking down the words in media reports to determine sentiment based on the part of speech and the intensity of the tone. For example, Tetlock measures sentiment by calculating the ratio of positive or negative words to the total number of words in media reports [8]. The second method uses the volume of media coverage as an indicator of sentiment. Fang and Peress, for instance, utilize the quantity of news reports from reputable newspapers as a variable for measuring media sentiment [9]. Fan et al. proposed the factor-augmented regularized model for prediction (FarmPredict) model, which combines both methods. This framework extracts data from financial news on Sina Finance(https://finance.sina.com.cn), including time, text, and stock information [10]. It summarizes each news article by counting words and evaluates sentiment scores for the period from 2015 to 2019. In their approach, all articles are treated as a collection of words, with sentiment words directly determining the sentiment score of the article. The model selects 10,000 common words from 1.1 million words for machine learning, filters out articles with similar content and publication times, and matches articles with stocks. Articles involving zero or multiple stocks are removed, and a down-sampling strategy is used to balance the number of articles. Hyperparameters are tuned during training, and sentiment words are further selected using penalized logistic regression (see Figure 4). However, the FarmPredict model proposed by Fan et al. only considers text data and does not address the impact of image data. Obaid and Pukthuanthong mention that images can shift attention from text, particularly when they convey strong emotions. They propose using the Google Inception v3 model for analyzing sentiment from image data. Although this model is widely used in practice and research, its high computational cost makes it challenging for market prediction applications [11].

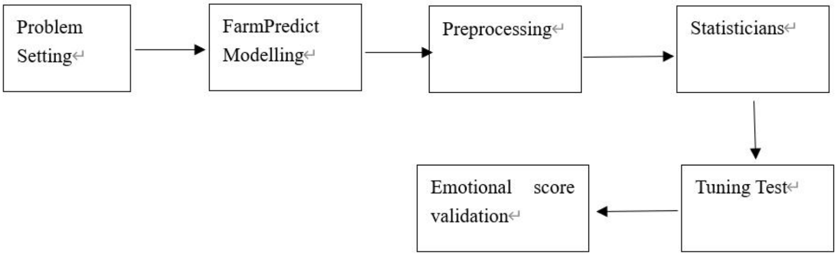

Fig. 4 Specific processing method of FarmPredict

3.2 Analyzing Investor Sentiment Through Social Media

As shown in Figure 5, similar to traditional news media analysis methods, after extracting a large amount of valid text data such as posts and tweets using web crawlers, researchers first organize and classify the text data. They then use sentiment dictionaries to categorize the semantic orientation of terms in the text. By weighting and summing the sentiment values from the dictionary for each term in a tweet, they obtain the overall sentiment of that tweet. Finally, the sentiment of all tweets from a given day is processed in various ways to determine the investor sentiment for that day. Different processing methods can lead to varying results in the relationship between investor sentiment and different economic indicators in the stock market, such as returns, asset prices, and trading volume. Generally, the sentiment of terms is categorized into positive, negative, and neutral. Littman et al. use the Semantic

Orientation Pointwise Mutual Information (SO-PMI) algorithm, which involves defining a set of positive and negative benchmark words with clear emotional orientations. The PMI (pointwise mutual information) values between candidate words and these benchmark words are calculated to measure the strength of association with the two sentiment benchmarks, and the sentiment orientation SO value is used to classify the words accordingly [12].

Clearly, the sentiment inference of candidate words is significantly influenced by the word frequency of individual benchmark words. To reduce errors caused by the uneven sample distribution of benchmark word sets, the PMI calculation formula needs to be standardized. Additionally, during sentiment inference, since candidate words may not have equal association strength with positive and negative words, setting a positive threshold can reduce the sensitivity of potential neutral candidate words to polarity sentiment, thereby improving the accuracy of the algorithm. Therefore, Huang and Yang et al. introduced standardization and the "Laplace correction" method, and based on this, set sentiment classification thresholds to propose an improved SO-LNPMI sentiment classification algorithm [13].

Many binary sentiment analysis models, such as the SO-LNPMI sentiment classification algorithm, may overlook the rich and multidimensional nature of Human feelings. To capture more aspects of public sentiment, Bollan et al. extended a well-established psychological measurement tool, the Profile of Mood States (POMS), to develop a new sentiment tracking tool called the Google-Profile of Mood States (GPOMS). This tool gauges emotions through six different dimensions: Calmness, Alertness, Confidence, Sense of Importance, Kindness, and Happiness.To adapt it for Twitter sentiment analysis, researchers expanded the original 72 POMS terms into a comprehensive dictionary containing 964 relevant terms. This expansion was accomplished by examining around 1 trillion word tokens that Google analyzed from publicly available web pages in 2006, focusing on co-occurrence patterns in 2.5 billion 4-grams and 5-grams. The expanded dictionary enables GPOMS to encompass a wider variety of naturally occurring emotional terms in tweets and align them with the corresponding POMS emotional dimensions [14]. Consequently, scores for each POMS emotional dimension are calculated as the weighted sum of co-occurrence values for each term in the GPOMS dictionary that matches terms in each tweet. Public investor sentiment for a given day is measured by summing the six-dimensional GPOMS emotional vectors for all tweets of that day and dividing by the total number of tweets. This results in a smooth time series, and the correlation with stock return impacts will be tested in subsequent analyses.

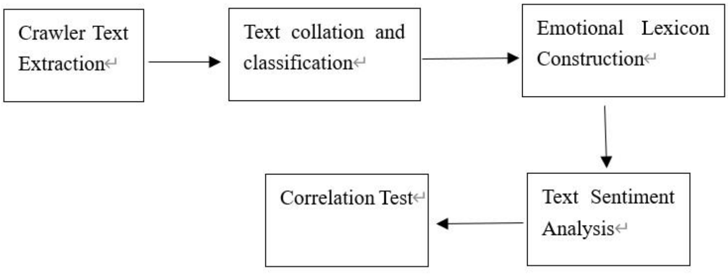

Fig. 5 The specific processing methods of social media sentiment analysis algorithms

FarmPredict Return Prediction Capability Test

Given that the consistency of our sentiment lexicon with sentiment has already been validated, it is crucial to assess the feasibility of using the FarmPredict algorithm to predict stock returns. In Fan et al.'s study, the prediction capability was tested by analyzing the correlation between beta-adjusted returns of relevant stocks and the calculated sentiment scores. Their method involved regressing panel data of beta-adjusted stock returns from January 2005 to December 2019, with control variables including stock size, purchase price, annual return beta coefficients, and alpha coefficients. To account for endogeneity, a one-week lagged return variable was also included. The relationship between sentiment scores and returns adjusted using beta returns is shown in Table 1. Only the results controlling for lagged returns and other stock-level variables are presented here. Despite the coefficients being smaller compared to those before control, the positive correlation between beta-adjusted returns and sentiment scores remains highly significant.

Table 1. Relationship between sentiment score and stock returns adjusted for beta

	FarmPredict	SESTM-Logistic	SESTM
sentiment	0.350	0.852	4.321
Lagged returns	×	×	×
Control variables	×	×	×
Earning Suprises	×	×	×
Time fixed effect	√	√	√
Adjusted	0.007	-0.001	-0.0001

Correlation Test of the SO-LNPMI Sentiment Classification Algorithm and the GPOMS Algorithm with the Stock Market

Regardless of whether using the SO-LNPMI sentiment classification algorithm or the GPOMS algorithm, the generated sentiment time series need to be tested for correlation with the stock market. For the SO-LNPMI sentiment classification algorithm, Huang and Yang et al. established a leading-lag relationship model of time series using the Granger causality test method. They conducted empirical analysis by testing different lag lengths and used the AIC criterion to select the optimal lag order. For the GPOMS algorithm, Granger causality analysis can also be used to examine the correlation between different dimensions of sentiment and economic indicators of the stock market. However, since Granger causality analysis relies on linear regression, and the connection between public sentiment and stock market value is almost certainly nonlinear, researchers compared the effectiveness of a self-organizing fuzzy neural network (SOFNN) model with two sets of input data predicting the Dow Jones Industrial Average (DJIA) values. One set of inputs included the DJIA values from the past three days, while the other set added the sentiment time series to improve prediction accuracy. This comparison assessed the impact of incorporating public sentiment information on the accuracy of prediction models and further demonstrated that including sentiment information significantly improves the precision of even the simplest models in forecasting DJIA closing values.

4 Conclusion

From the research on investor sentiment and the stock market both domestically and internationally, differences in conclusions are evident. In domestic stock markets, which have not been established for long, individual investors are more susceptible to herd behavior, making stock prices more responsive to investor sentiment. This effect is most pronounced in smaller, speculative stocks. In financial markets, traditional media news can significantly influence individual investors' decisions through the investment tendencies and emotions it reveals. Meanwhile, with the advancement of technology, social media platforms are rapidly developing. An increasing number of individual investors gather on financial forums to extract information and share their opinions, forming a substantial public sentiment.

This paper, on one hand, focuses on sentiment analysis from traditional media news. It introduces two approaches for sentiment analysis and describes a specific algorithm model, FarmPredict, which combines both methods. Through empirical analysis, it has been validated that this model achieves excess returns in practical applications. On the other hand, from the perspective of social media sentiment analysis, this paper introduces the SO-LNPMI sentiment classification algorithm and the GPOMS algorithm. The correlation between these algorithms and the stock market is tested using Granger analysis and the self-organizing fuzzy neural network (SOFNN) model. The Granger analysis of the SO-LNPMI sentiment classification algorithm reveals that the positive sentiment described by this algorithm is a Granger cause of the Shanghai Composite Index return, while the negative sentiment is not a Granger cause of the return but is a Granger cause of the trading volume. The Granger causality between negative sentiment and trading volume is more pronounced than that between positive sentiment and trading volume. Adding the investor sentiment time series derived from the GPOMS algorithm to the SOFNN model and improving the model's prediction accuracy further demonstrates the relevance of investor sentiment to the stock market.

As media and lifestyle habits evolve, more people are inclined to browse images and videos, and traditional text-based news media is gradually shifting towards short videos and multimedia content. Therefore, future research on alternative data for predicting stock returns should incorporate dynamic data capture from images and videos for a more comprehensive analysis of investor sentiment. In practice, social media users' posting patterns are highly variable. Improving natural language processing techniques and integrating them with investor sentiment measurement is a future research direction. Additionally, exploring the interactive relationship between different investor sentiments and the stock market across various stages of economic cycles is another important area for future study.

Author Contributions

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

References

1. DeLong J. B., Shleifer A., Summers L. H., et al.: Noise Trader Risk in Financial Markets. The Journal of Political Economy **98** (4), 703-738 (1990).
2. Kumar A., Lee C.: Retail investor sentiment and return co-movements. The Journal of Finance **61** (5), 2451-2486(2006).
3. Wurgler J. A., Baker M. P.: Investor Sentiment and the Cross-Section of Stock Returns. Economic Management Journal **61** (4), 1645-1680 (2006).
4. Sankaraguruswamy S., Mian G. M.: Investor Sentiment and Stock Market Response to Corporate News. SSRN Electronic Journal, (2008).
5. Song D., Chi L., Zhuang X.: Investor sentiment in the Chinese stock market: an empirical analysis. Applied economics letters **19** (4), 345-348 (2012).
6. Zhang B., Hu W. X., Yang X.: Research on the Mechanism of Investor Sentiment Generation. China Management Science **29** (01), 185-195 (2021).
7. Xin Y.: Analysis of the Impact of Investor Sentiment on Stock Returns. Hebei University of Finance, (2023).
8. Tetlock, Paul C.: Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance **62** (3),1139-1168 (2007).
9. Fang L., Peress J.: Medie Coverage and the Cross-section of Dtock Returns. The Journal of Finance **64** (05), 2023-2052 (2009).
10. Fan, J., Xue L. and Zhou Y., 2021.: How much can machines learn finance from Chinese text data?. Management Science, (2024).
11. Obaid K., Pukthuanthong K.: A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. Journal of Financial Economics **144** (1),273-297 (2021).
12. Turney P. D., Littman M. L.: Measuring Praise and Criticism: Inference of Semantic Orientation from Association. Acm Transactions on Information Systems **21**(4), 315-346 (2003).
13. Huang C. X., Wen S. G., Yang X., et al.: Research on the Interactive Relationship Between Individual Investor Sentiment and Stock Price Behavior. China Management Science **28** (3), 191-200 (2020).
14. Bollen J., Mao H., Zeng X.: Twitter mood predicts the stock market. Journal of computational science **2** (1),1-8 (2011).