Investor Sentiment and Stock Market Performance in the Big Data Environment

Zhihan Wang[1,a,*]

[1]*School of Accounting, Southwest University of Finance and Economics, 555 Liutai Avenue, Chengdu, China*
a. zhihanwang.swufe@gmail.com
**corresponding author*

Abstract: This study investigates the influence of social media sentiment on stock market performance by analyzing investor comments on the Snowball platform for Yunnan Baiyao (00538.SZ) and CICC (601995.SH) (China International Capital Corporation) between July and September 2024. Using web-scraping technology and sentiment analysis, the research categorizes comments into positive, neutral, or negative, applying regression models to investigate the relationship between the number of posts reflecting investor sentiment and stock trading performance, while controlling for variables such as company assets, P/E ratio, and market turnover rate. The results reveal that both positive and negative sentiments significantly affect stock trading behavior, with negative sentiment having a stronger impact. Additionally, company size and market fundamentals consistently influence trading volume and turnover rate. The study concludes by recommending that investors rely more on fundamental analysis rather than short-term sentiment fluctuations to achieve long-term stability in their investments.

Keywords: investor sentiment, stock performance, data scraping.

1. Introduction

With the development of information technology, social media has become an integral part in people's lives. In the financial market, investors increasingly rely on social media platforms to obtain information and share their views, which makes social media an important factor influencing market sentiment and behavior. Compared to traditional media (e.g., newspaper news), people's comments on social media tend to better reflect their genuine emotions, as the anonymity afforded by the internet encourages people to express their true thoughts more freely. However, this anonymity also facilitates the manipulation of netizens' emotions, and the influence of social opinion is frequently underestimated. Especially in the process of investment decision-making, investors' emotions are easily influenced by others, following the trend and forming a herd effect.

Research on previous studies show that there is a significant correlation between social media sentiment and stock market performance. Many scholars have explored this relationship through various methods, and the studies show that investor sentiment can significantly affect the volatility of the stock market, the rate of return, and other market indicators. However, different studies have used different methods and models, and the results show that the path of sentiment's influence on the market is complex and diverse, and its predictive effect is affected by a variety of factors.

Snowball Platform is a social investment platform focused on providing professional financial data, market analysis and interactive discussions for individual and institutional investors. The platform not only covers a wide range of financial products such as stocks, funds, bonds, etc., but also provides users with real-time market quotes, company announcements, financial statements and other comprehensive financial information to help investors conduct in-depth market analysis and decision-making.

The core feature of Snowball is its user-generated content (UGC) model, which allows users to form a high-quality interactive community by posting investment insights, market analyses, comments and other information. This provides investors with diverse perspectives and facilitates a more comprehensive understanding of market movements. In addition, the platform provides a simulated investment function, enabling users to rehearse their investment strategies and gain experience before investing.

This study uses web crawler technology to collect investor stock comments from Snowball platform and rate the sentiment of the collected comments through a sentiment thesaurus. Based on the rating results, the posts are classified into positive, negative, and neutral sentiments. Regression models are constructed to analyze the relationship between social media sentiments and stock market performance by controlling variables. This study can not only provide new empirical evidence for understanding the market effect of social media sentiment, but also provide decision-making references for investors to help people better cope with market volatility.

2. Literature Review

In recent years, the exploration of the relationship between social media sentiment and stock market performance has garnered significant attention in academia. Various studies have delved into different aspects and perspectives within this domain, providing valuable insights into the complex interplay between the two.

Firstly, a series of research has focused on the direct impact of social media sentiment on the stock market. Early work by Bu suggested that while social media sentiment does influence the stock market, its predictive power is relatively limited. [1] Xu and Tian utilized text mining and random forest methods to discover the volatility spillover effect of the internet fund market on the stock market, emphasizing that the impact is actually more pronounced and economically significant.[2] Collectively, these studies indicate that there is a certain correlation between social media sentiment and stock market performance, but assessments of its predictive ability vary.

Furthermore, researchers have delved deeper into the specific mechanisms through which social media sentiment exerts its influence. Du and Lv found that positive social media posts generally have a positive impact on the stock market, while negative posts have a negative impact. [3] This finding echoes the results of Feng and Zhang, who further demonstrated that social media sentiment can drive stock market volatility, with positive sentiment potentially triggering market upswings and negative sentiment leading to market downturns.[4] These studies not only confirm the direct impact of social media sentiment on the stock market but also reveal its potential to drive market fluctuations.

On the methodological front, researchers have employed various analytical methods and models to capture the link between social media sentiment and the stock market. Plenzick used ANEW sentiment scores to analyze Twitter data, attempting to predict stock market trends. [5] Wang, leveraging the Twitter API and R language for sentiment analysis, found that incorporating sentiment indicators and nonlinear models (such as NARX) can effectively predict stock market dynamics. [6] Du used simple Bayesian models and Granger causality tests to explore the relationship between investor sentiment and short-term market volatility. [3]

In summary, these studies collectively demonstrate that social media sentiment has a significant impact on stock market performance, and this impact is not confined by geographical boundaries.

Researchers, using different analytical methods and models, have successfully captured important links between sentiment and market volatility, providing strong support for our deeper understanding of the role of social media sentiment in financial markets.

3. Research Design

3.1. Research Sample

This study selects comments on the Snowball platform from July 4th to September 5th, 2024, focusing on two company stocks in China's A-stock market, Yunnan Baiyao (00538.SZ) and CICC (601995.SH) as the research objects. These two stocks are highly representative due to their relatively active yet extensive discussions on the Snowball platform. Yunnan Baiyao, a leading company in China's traditional medicine sector, showed significant volatility in its stock price during the observation period of July 19-September 5, 2024, which reflects the changing dynamics of China's pharmaceutical market. Meanwhile, the stock price of CICC, a prominent leader in the financial services sector, is also profoundly influenced by fluctuations in the global financial market and changes in policy regulation. The observation period for this stock review spans from July 4th to September 5th, 2024.By analyzing these two stocks, we can not only capture variations in market sentiment across different sectors but also gain access to a vast amount of commentary data on the Snowball platform for in-depth sentiment analysis, thanks to the intense discussions surrounding them. The diverse nature of these two stocks during the study period provides an ideal sample base for analyzing the relationship between investor sentiment and stock price volatility, enhancing the comprehensiveness and convincingness of our study.

3.2. Data collection method

This study will use web crawler technology to obtain data on investor comments under two stocks, Yunnan Baiyao (00538.SZ) and CICC (601995.SH), from the Snowball platform. The study will focus on collecting basic information such as the content and posting time of the comments to capture investors' sentiment expressions under different market conditions. The author first analyzed and labeled the sentiment characteristics of a sample of vocabulary. For example, words indicating positive sentiments include "rise," "buy-in," "big profit," "set new highs," etc.; words indicating negative sentiments include "plummet," "sell," "loss," "cancel order," "blow-up," "panic," "be deceived," etc. The remaining words, which do not carry obvious sentiment characteristics, are labeled as neutral words. Then, using Snow NLP in Python, text analysis and scoring of these comments are conducted, with a score range of 0-1. Based on the scoring results, the comments are categorized into three categories: positive, neutral and negative, where a score greater than 0.5 is positive, a score equal to 0.5 is neutral, and a score less than 0.5 is negative. Since this paper focuses on the impact of investors' positive and negative emotions on their investment behavior, this study selects the total number of positive emotional comments and the total number of negative emotional comments on each stock trading day as the explanatory variables. And the daily trading volume and turnover rate of the individual stocks of these two stocks are recorded.

3.3. Variable analysis

This study constructs a comprehensive set of variables to analyze the pool based on the collected data plus control variables. The daily number of total, positive and negative comments derived from the sentiment analysis served as explanatory variables. The logarithm of daily individual stock trading volume and the turnover rate are used as explained variables. Simultaneously, the logarithm of the company's total assets, the price-earnings ratio, the logarithm of the company's total market capitalization, and the market turnover rate are considered as control variables

3.4. Model Construction

This study will use the explained variables (ln_Volume, Turnover_Rate), the explanatory variables (Post, Positive_Post, Negative_Post) and the control variables (ln_Assets, P/E,ln_Market_value, Market_turnover) to construct a regression analysis model. The model takes the form:

Explained variables =α Explanatory variables +βControl variables +ε.

For example,

In_Volume=αPostive_Post+β1 ln_Assets+β2P/E+β3ln_market_value+β4Market_turnover+ε

4. Empirical regression results

4.1. Descriptive statistics

Table 1: Results of Descriptive Statistics.

Variables	N	Max	Mean	Min	Standard deviance
Volume	81	274289.0000	87039.9346	38450.3000	41715.3382
Turnover	81	0.9660	0.3752	0.1315	0.1795
Post	81	180.0000	21.6543	5.0000	23.3889
Positive_post	81	83.0000	10.2173913	2.0000	11.8264
Negative post	81	62.0000	7.7826	1.0000	9.0647
Asset	81	613694656004	373093289685.38	56874351095.19	275830942210.30
P/E	81	24.2462	22.5107	20.9365	0.77030048
Market_value	81	13801127.4000	10142687.6000	8817825.78000	1541440.4400
Market_turnover	81	2.2900	1.0457	0.4681	0.5238

As shown in Table 1 of the descriptive statistics analysis, the average value of Volume is 87039.93, indicating significant fluctuations in the trading volume of stocks within the sample. The minimum value is 38450.3, the maximum value is 274289, and the standard deviation is 41715.34, highlighting substantial differences in trading volumes among individual stocks. The average Turnover is 0.3752, with a standard deviation of 0.1795, and the minimum and maximum values are 0.1315 and 0.966, respectively, suggesting a relatively stable liquidity of stocks in the market. The average Price/Earnings Ratio (P/E) is 22.51, with a small standard deviation and relatively stable fluctuations, indicating a consistent valuation of companies within the sample. The average Market Turnover is 1.0457, suggesting a high level of market activity. The average value of the Post variable is 21.6543, with a standard deviation of 23.3889, indicating a relatively dispersed distribution of this variable within the sample, with some degree of extreme values. These statistical results provide preliminary data characteristics for subsequent analysis, revealing the basic distribution of variables within the sample.

4.2. OLS empirical regression results analysis

Table 2: Empirical Regression Results between Total Number of Posts and Stock Performance.

	(1) ln_Volume	(2) Turnover
Post	0.00461***	0.00166**

Table 2: (continued).

	(3.18)	(2.58)
ln_Assets	26.24*	14.76***
	(1.91)	(3.10)
P/E	6.886*	3.881***
	(1.95)	(3.16)
ln_Market_Value	-150.0*	-85.33***
	(-1.87)	(-3.07)
Market_turnover	0.910***	0.343***
	(4.52)	(4.26)
_cons	1610.7*	915.3***
	(1.86)	(3.06)
N	81	81
R2	0.4104	0.5105

t statistics in parentheses
*p< 0.1, **p< 0.05, ***p< 0.01

This paper employs the Ordinary Least Squares (OLS) regression method to test the empirical model. The regression results presented in Table 2 uncover crucial relationships between the total number of daily posts and daily trading indicators of stocks. In Model (1), the total number of posts (*Post*) is significantly positively correlated with stock trading volume (*ln_Volume*) at the 1% level, suggesting that a higher number of posts about a particular stock on a given day is associated with increased trading volume of that stock in the same period. In Model (2), the total number of posts (*Post*) also exhibits a significant positive correlation with turnover rate (*Turnover*), implying that a greater number of discussion posts about a stock on a specific day leads to a higher turnover rate for that stock. The total number of posts reflects investor sentiment, indicating that in the big data era, when a stock attracts more investor attention, its trading performance immediately mirrors the sentiment of market investors.

Additionally, regarding control variables, the logarithm of assets (*ln_Assets*) has a significant positive impact on both trading volume and turnover rate, suggesting that companies with larger asset sizes also experience higher trading volumes and turnover rates. The Price-to-Earnings (*P/E*) ratio similarly contributes positively to trading volume and turnover rate, implying that companies with higher P/E ratios also exhibit higher trading volumes and turnover rates. However, the logarithm of total market value (*ln_Market_value*) negatively influences trading volume and turnover rate, indicating that companies with larger market capitalizations often have lower trading volumes and turnover rates. Furthermore, there is a significant positive relationship between market turnover (*Market_turnover*) and individual stock's trading volume and turnover rate.

Table 3: Empirical Regression Results between Total Number of Positive Posts and Stock Performance.

	(1)	(1)
	ln_Volume	Turnover
Postive_post	0.00845**	0.00306*
	(2.28)	(1.92)
ln_Assets	25.98*	14.66***
	(1.86)	(3.02)
P/E	6.815*	3.854***

Table 3: (continued).

	(1.90)	(3.08)
ln_Market_value	-148.5*	-84.75***
	(-1.82)	(-3.00)
Market_turnover	0.912***	0.344***
	(4.55)	(4.25)
_cons	1593.6*	908.9***
	(1.81)	(2.98)
N	81	81
R2	0.4122	0.5148

t statistics in parentheses
* p < 0.1, ** p < 0.05, *** p < 0.01

The regression results in Table 3 reveal the crucial relationship between the daily total of positive posts and stock trading indicators. In Model (1), the daily total of positive posts is significantly positively correlated with the current stock trading volume (ln_Volume) at the 1% level, indicating that the higher the total number of positive posts about a stock on a given day, the higher its turnover rate. In Model (2), the daily total of positive posts is also significantly positively correlated with the current turnover rate (Turnover) of the stock, meaning that the more discussion posts about a stock on a given day, the higher its turnover rate. The total number of posts reflects investor sentiment, suggesting that in a big data environment, when investors are optimistic about a stock, they actively express their sentiment online, and when a stock attracts more investor attention, its trading performance immediately reflects the market's investor sentiment.

Table 4: Empirical Regression Results between Total Number of Negative Posts and Stock Performance.

	(1) ln_Volume	(2) Turnover
Negative_post	0.0139***	0.00515***
	(4.06)	(3.11)
ln_Assets	25.61*	14.51***
	(1.91)	(3.14)
P/E	6.730*	3.819***
	(1.96)	(3.21)
ln_Market_value	-146.4*	-83.91***
	(-1.87)	(-3.12)
Market_turnover	0.901***	0.340***
	(4.46)	(4.26)
_cons	1572.3*	900.1***
	(1.86)	(3.10)
N	81	81
R2	0.4074	0.5091

t statistics in parentheses
* p < 0.1, ** p < 0.05, *** p < 0.01

The regression results in Table 4 uncover the crucial relationship between the total number of negative posts and stock trading volume indicators. In Model (1), the daily total number of negative posts is significantly positively correlated with stock trading volume at the 1% level, indicating that the higher the total number of negative sentiment posts about a stock on a given day, the higher its trading volume. In Model (2), the daily total number of negative posts is also significantly positively correlated with stock turnover, suggesting that the more negative discussion posts a stock receives, the higher its turnover. The total number of negative posts reflects investor sentiment, and these results demonstrate that in a big data environment, when a stock faces increased negative attention, its performance will quickly adjust to reflect the negative sentiment expressed by investors.

5. Conclusion and Recommendation

Utilizing web scraping and data mining techniques, this paper systematically collected textual data from stock commentaries on Snowball Forum, and subsequently conducted meticulous sentiment classification on these texts. The research aims to delve into the intrinsic correlation between investor sentiment and stock trading performance, aligning closely with the rapid development of big data technology in the current Internet era and its increasingly profound integration into people's daily lives.

By synthesizing the results of six regression analyses, this study reveals that investor sentiment has a significant impact on both stock trading volume and turnover rate. Specifically, both positive and negative comments exert an influence on trading volume and turnover rate, with negative sentiment having a particularly prominent effect on stock trading volume in certain models. Furthermore, total assets and price-to-earnings ratio exhibit stable positive effects across various regression models, suggesting that companies with larger total assets and higher price-to-earnings ratios tend to have higher trading volumes and turnover rates. At the same time, market capitalization demonstrates a negative effect, indicating that as a company's market capitalization increases, its turnover rate decreases, which reflects the stability characteristic of large-cap stock trading to some extent. Notably, the overall market turnover rate has a significant positive influence on individual stock turnover rate, indicating that the overall market activity can effectively drive trading activities of individual stocks. Overall, through empirical analysis, this study provides strong evidence for the significant influence of investor sentiment, corporate characteristics, and market factors on stock trading behavior, offering solid theoretical support for a deeper understanding of investor behavior and market reactions in the stock market.

Based on our analysis of market variables and investor sentiment’s effects on trading behavior, it is evident that both positive posts and fundamental market indicators have a significant correlation with trading volume and turnover rate. This finding underscores the role of investor sentiment in influencing stock market dynamics, as it often serves as a proxy for collective investor mood and anticipation regarding future performance. However, we must consider that emotional information often leads to volatility and short-term fluctuations, which can distort true asset value in the market. Investors who heavily rely on sentiment-driven information may react prematurely to market changes, basing their decisions on transient shifts in mood rather than underlying financial health or performance indicators.

This tendency can lead to impulsive or irrational investment decisions, as market sentiment may not fully capture or reflect long-term fundamentals, creating discrepancies between a stock’s intrinsic value and its market price. Consequently, while investor sentiment is crucial, it should be considered alongside traditional fundamental analysis. A balanced approach that combines sentiment insight with a thorough evaluation of financial health and market trends can empower investors to make more informed, resilient decisions, avoiding the pitfalls of short-term speculation and aiming instead for sustainable, long-term returns.

Therefore, it is recommended that investors should remain rational when making investment decisions and avoid trading solely based on short-term emotional fluctuations on social media. While market sentiment on social platforms can sometimes offer valuable insights, relying heavily on it without considering underlying factors can lead to misguided judgments and reactionary moves. Instead, they should comprehensively assess investment risks and returns, considering a range of indicators such as company fundamentals, industry trends, and long-term market performance. By doing so, investors create a more stable foundation for their portfolios, better positioning themselves to withstand market volatility.

At the same time, it is important to be alert to the noise in the market, which can be especially misleading when the market fluctuates dramatically. The power of mass sentiment can amplify swings, often causing sudden spikes or drops that don't reflect the true value of assets. In such moments, it becomes crucial not to blindly follow the trend or be swayed by emotions, as doing so can lead to speculative losses rather than informed gains. Maintaining a calm and rational mindset enables investors to act based on analysis rather than impulse, enhancing the potential to obtain sound returns in long-term investment. This balanced approach helps investors capitalize on genuine opportunities while safeguarding their strategies from the unpredictable tides of sentiment-driven trading.

It is hoped that investors will pay more attention to fundamental analysis and maintain an appropriate distance from market emotions in their future investment decisions, so as to avoid short-term fluctuations affecting their long-term investment objectives.

References

[1] *Bu, H., Xie, J. F., Li, J. H., & Wu, J. J. The impact of investor sentiment based on stock reviews on the stock market. Journal of Management Science, 2018(Vol.21,No.4).*

[2] *Xu, C. M., & Tian, J. Q. The impact of internet funds on the stock market. Journal of Nanjing Audit University, 2016(6).*

[3] *Du, W. A., &Lv, J. L. Social media big data, investor sentiment, and IPO underpricing. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 2018(Vol.20,No.3).*

[4] *Yao, J. Q., Feng, X. J., Wang, Z. J., Ji, R. J., & Zhang, W. Language, emotion, and market influence: Based on a financial sentiment dictionary. Journal of Management Science, 2021(Vol.24, No.5).*

[5] *Plenzick, C. Can You Really Predict Markets with Twitter? 2016.*

[6] *Wang, Y. A big data study on the sentiment analysis of social networks and nonlinear system modelling. 2018.*