

The Relationship Between Investor Sentiment and Stock Returns under Big Data

Jincheng Dai^{1,a*}

¹*NOIC Academy, 50 Featherstone Avenue, Canada*

a. 2749796011@qq.com

**corresponding author*

Abstract: In this age, online search behavior has become the key to reflecting investors' attention and decision-making, and big data technology has unveiled a tight correlation between search volume and stock market trends, highlighting its potential as a real-time indicator of investor sentiment. This paper empirically analyzes the relationship between investor sentiment and stock returns, using the Internet search index as a proxy variable for investor sentiment based on big data. Through a panel model regression test conducted on 466,663 sample data from 1,974 companies on China's A-share main board market across 242 trading days in 2023, the results reveal a statistical correlation between the Internet search index and stock market performance. Specifically, the Internet search index significantly affects the individual stock price fluctuation, individual stock volume, individual stock transaction amount and turnover rate at the 1% significance level, indicating that the higher the investor sentiment, the higher the stock's day's return. We hope that investors will keep their sanity when investing and not be easily motivated by emotions and the media and learn to think independently and objectively analyze the data to decide to invest in the project. Chasing up and down is not desirable and irrational which reflects the psychological characteristics of gamblers do not apply to today's market.

Keywords: Investor Sentiment, Stock Returns, Big Data

1. Introduction

In today's highly integrated and digitized era, online searches have become a critical bridge connecting information resources and user demands. Characterized by their customization capabilities, these searches are significantly driven by users' external environments and intrinsic needs. Particularly in the financial domain, investors' online search behaviors not only directly reflect their attention to stock market information but also profoundly shape their investment preferences and decision-making. The advent of big data technology has greatly enhanced the efficiency of capturing and analyzing these behaviors, allowing researchers to identify patterns and correlations between online searches and stock market trends. This underscores the close relationship between search volume and stock market indices, as well as its potential as a real-time indicator of investor sentiment.

Investor sentiment, first conceptualized by Stein [1], refers to systematic biases in investors' expectations of future markets. It reflects a composite of optimism, pessimism, and neutral attitudes, embodying the collective investment intentions and expectations of market participants. Investor

sentiment has a profound impact on stock markets, especially in the era of big data, which amplifies the continuous influence of information on retail investors' decision-making. In particular, online search behavior has become a critical expression of investor sentiment, as it reflects both individual reactions to market events and broader collective trends. In the Chinese stock market, investor behavior is shaped by several unique institutional factors, which not only influence sentiment but also manifest clearly in online search activity.

This paper highlights three critical factors that make investors more prone to emotional decision-making: China's centralized governance model, sensationalist media practices, and low levels of education and financial literacy among retail investors. These factors collectively drive herd behaviors, such as "chasing highs and selling lows," exacerbating market volatility, and undermining rational investment practices. Importantly, these same factors also amplify fluctuations in online search volumes and content trends, making search behavior an even more valuable proxy for investor sentiment analysis in the Chinese context.

The centralized governance model in China plays a pivotal role in shaping investor sentiment and decision-making. Government policies often exert direct and indirect control over market activities, significantly influencing investor expectations. Sudden policy shifts, such as restrictions on trading or unexpected fiscal measures, frequently disrupt market stability and trigger emotional responses among retail investors. For example, in 2021, the central government's restrictions on real estate purchases and loans redirected some capital into the stock market. Simultaneously, the high-frequency regulation of the capital market, including limitations on leveraged funds and intensified scrutiny of private equity funds, led to a decline in investor sentiment. This dual pressure from real estate restrictions and stock market policies resulted in a general lack of confidence, increased market volatility, and heightened emotional sensitivity. Such rapid and unpredictable changes in the regulatory environment are often closely followed by corresponding spikes in online searches related to policy terms, stock performance, and market trends.

The second factor is the sensationalism of media reporting. In China, the media comprises three main components: individual media, newspapers, and official media. The government controls official media and regulates individual media and newspapers through a mature system of centralized governance, often prioritizing emotional engagement over factual accuracy. Sensational headlines, exaggerated narratives, and biased commentary heavily influence investor perceptions and decision-making. Studies have shown that pessimistic media coverage can lead to panic-driven selloffs, while overly optimistic news may trigger speculative bubbles. In the Chinese context, where big data enables real-time analysis of investor behavior, the media's role in shaping collective sentiment is particularly pronounced. This impact is further evidenced in the changes in search volumes for keywords associated with trending media stories, making search data a critical tool for understanding how media sensationalism influences investor behavior.

The educational background and financial literacy levels of Chinese retail investors significantly impact their investment decisions. Research indicates that many investors lack a comprehensive understanding of market fundamentals, often making decisions based on short-term emotional reactions rather than long-term strategic analysis. This vulnerability is compounded by China's rapidly developing yet immature financial markets, where novice investors frequently rely on unverified information and exhibit herd behaviors. These tendencies are mirrored in their online search patterns, where spikes in certain keyword searches often align with periods of heightened market volatility, further highlighting the connection between investor sentiment and search behavior.

The three causes mentioned earlier have a pivotal impact on the direction of the Chinese stock market but are also reflected in the data. The stock market fluctuations are often emotionally influenced by the fluctuations in mood and changes in the rate of return by the three factors indirectly or directly, and the mood of the direct changes. With the wide application of big data technology, the

efficiency of information dissemination has been significantly improved. In this context, investors' online search behavior has become an important indicator reflecting changes in their sentiment. Therefore, this paper empirically examines the relationship between Internet search index and stock returns using a panel regression model using Internet search index as a measurement variable of investor sentiment with a sample of 242 trading days of Shanghai and Shenzhen A-share main panel companies in 2023. An in-depth investigation of the intrinsic connection between investor sentiment and Internet search behavior not only helps us better understand the deep-seated causes of stock market fluctuations but also provides investors with a more scientific basis for decision-making. At the same time, this study also provides useful references and insights for the government and regulators to formulate more effective market control policies and optimize investor education mechanisms. In conclusion, this topic has important theoretical value and practical significance and has a positive effect on promoting the sound development of the financial market.

2. Literature Review

In recent years, both domestic and international scholars have conducted extensive research on the relationship between investor sentiment and the stock market. These studies consistently highlight the critical role of investor sentiment in influencing stock returns. Brown and Cliff [2] explored investor sentiment, defining it as investors' expectations and attitudes toward the stock market, encompassing optimism, pessimism, and neutrality. They found that investor sentiment has predictive power in financial markets, significantly influencing stock price trends and volatility. Wurgler and Baker [3] analyzed stock cross-sectional returns and discovered that when overall investor sentiment is low, small-cap and high-volatility stocks often achieve higher subsequent returns. This indicates that emotional investors tend to chase popular stocks, contributing to price reversals and revealing the impact of sentiment on market pricing efficiency.

In China, researchers such as Wen et al. [4] categorized investor sentiment into positive and negative emotions and studied their effects on returns. Their findings indicated that positive sentiment significantly influences returns, while the impact of negative sentiment is less pronounced. Wang Chun and Wang Jin Men [5] confirmed that investor attention significantly affects stock returns and trading volumes. They noted that investor attention often amplifies cognitive biases among retail investors, triggering excessive trading behavior and influencing turnover rates. Thus, investor attention affects not only stock prices but also market activity and liquidity.

Search indices, as proxies for investor sentiment, have been widely applied in related studies. Da et al. [6] pioneered the use of keyword search volumes as proxies for investor attention and analyzed the Russell 3000 Index using Google Search Volume Index (SVI). They found that keyword search volumes directly measure investor attention, highlighting the close connection between investor sentiment and stock markets. Liu [7] leveraged data from Baidu Index, 360 Index, and He Xun Index to explore how the timing and heterogeneity of investor attention affect stock prices. Khan et al. [8] used Google Trends data to construct sentiment indices, uncovering a significant causal relationship between sentiment and overall stock indices, further confirming the critical role of sentiment in the stock market.

In conclusion, scholars have comprehensively examined the relationship between investor sentiment and the stock market using diverse research methods, rich data sources, and unique analytical perspectives. Investor sentiment significantly influences stock returns through multiple complex mechanisms. First, sentiment directly impacts stock prices through representation mechanisms: positive sentiment often signals market optimism, driving up prices, while negative sentiment may lead to cautious market behavior, suppressing prices. Second, sentiment affects investors' decision-making and behavioral patterns, with heightened market sentiment often triggering herd behaviors that exacerbate price volatility. Lastly, sentiment influences information

processing mechanisms, shaping how investors interpret and react to market data. Emotional investors may amplify positive information while overlooking negative signals, leading to deviations from fundamental values and challenging market efficiency.

3. Data, Variables and Methodology

3.1. Data Sources

The investor sentiment data and stock market-related data utilized in this study are sourced from the China Research Data Service Platform (CNRDS). The sample period covers all trading days in 2023, spanning from January 3 to December 29, totaling 242 trading days. The sample includes all A-share main board companies listed in Shanghai and Shenzhen, amounting to 1,974 firms and comprising 466,663 valid sample data points.

3.2. Variable Selection

3.2.1. Investor Sentiment

Following the approaches of Da et al. [6] and Khan et al. [8], this study adopts the Internet search index as a proxy for investor sentiment. The data is derived from the CNRDS Web Search Volume Index (WSVI) database. This database integrates various web search indices and news sentiment data to reflect public sentiment and the popularity of listed companies. It serves as a key indicator of company visibility and its changes over time. The search index (SVI), calculated as the aggregated search values for keywords such as stock codes, company abbreviations, and full company names, is used as the independent variable in this study.

3.2.2. Dependent Variables

To comprehensively explore the impact of investor sentiment on stock return performance, this study selects four variables as dependent variables: daily stock price fluctuations (*chg*), daily trading volume (*daily_volume*), daily transaction amount (*daily_turnover*), and turnover rate (*turnover_rate*) of individual stocks.

3.2.3. Control Variables

This paper employs a panel model for regression and controls for individual and time effects. Consequently, two variables with intraday variability are adopted as control variables in this paper, namely, the daily total market value of individual stocks (*market_value*) and the daily percentage change in the trading market (*mk_chg*) where the individual stocks are listed.

All variable symbols and their corresponding definitions are shown in Table 1.

Table 1: Variables and Definitions

Nature of the Variable	Variable names	Variable Definition
Dependent Variable	<i>chg</i>	Daily price fluctuation of the individual stock (unit: %)
	<i>daily_volume</i>	Daily volume of the individual stock (unit: share)
	<i>daily_turnover</i>	Transaction amount of the individual stock (unit: yuan)
	<i>turnover_rate</i>	Turnover Rate of the individual stocks (unit: %)

Table 1: (continued).

Independent Variable	<i>SVI</i>	Search Index
Control Variable	<i>market_value</i>	Current total market capitalization of the individual stocks(unit: yuan)
	<i>mk_chg</i>	Daily market percentage change(unit: %)

3.3. Methodology

This paper employs a panel data model for empirical regression analysis, clustered at the firm level while controlling for time-fixed effects. Due to the large numerical scale of variables such as trading volume, transaction amount, search index, and market capitalization, these variables are log-transformed using the natural logarithm to ensure consistency in the regression analysis. They will be denoted as *ln*daily_volume, *ln*daily_turnover, *ln*SVI, and *ln*market_value, respectively.

Taking daily stock price fluctuations as the dependent variable as an example, the specific form of the panel regression model is as follows:

$$Chg_{i,t} = \alpha_{i,t} + \beta_1 \ln SVI_{i,t} + \beta_2 \ln market_value_{i,t} + \beta_3 mk_chg_{i,t} + \varepsilon_{i,t}$$

4. Empirical results

4.1. Descriptive Statistics

The descriptive statistics are presented in Table 2.

Table 2: Descriptive Statistics

stats	max	p50	min	mean	sd
<i>chg</i>	221.5500	-0.0400	-25.6700	0.0238	2.2575
<i>daily_volume</i>	2400000000	8440542	2400	19900000	43600000
<i>daily_turnover</i>	19200000000	74900000	21552	200000000	440000000
<i>turnover_rate</i>	85.0058	0.9144	0.0006	1.9071	3.5125
<i>SVI</i>	387614	808	0	1335	2283
<i>market_value</i>	2400000000000	8150000000	281000000	30500000000	116000000000
<i>mk_chg</i>	2.7100	-0.0400	-2.2800	-0.0250	0.7919

The fluctuation range of the stock's change percentage (*chg*) is remarkable, with a maximum of 221.55%, while the median value (P50) is only -0.04%, indicating extreme shifts in market sentiment. The minimum value drops to -25.67%, further underlining the volatility of the market. Overall, the average change percentage is 0.0238%, with a standard deviation of 2.2575%, revealing a significant degree of dispersion.

The internet search index (SVI), as an important indicator measuring investors' attention and sentiment, spikes to a maximum of 387,614, reflecting extremely high investor focus during specific periods. The median value (P50) is 808, indicating the normal level of attention. The minimum value is 0, showing that investors may completely ignore the stock at certain times. The average value is

1,335, with a standard deviation as high as 2,283, highlighting the substantial fluctuations in investors' attention.

At the transactional level, the average daily volume (*daily_volume*) of the stock reaches 19.9 million shares, demonstrating a certain level of market activity. The average transaction amount (*daily_turnover*) is 200 million yuan, reflecting the scale of capital flows in the market. The average turnover rate (*turnover_rate*) is 1.9071%, revealing the relative frequency of market transactions.

In addition, the average market value (*market_value*) of individual stocks for the day is 30.5 billion yuan, showing the overall value scale of the market. The average market change percentage (*mk_chg*) is -0.025%, reflecting a slight downward trend in the market's overall performance.

4.2. Panel Regression Statistics

This paper uses a panel model to regress the data and controls for individual firm effects as well as time effects. Table 3 presents the results of the empirical regression tests conducted on the entire sample. Columns (1) (2) (3) and (4) of Table 3 show the regression results when the dependent variables are individual stock price fluctuation (*chg*), individual stock volume (*Indaily_volume*), individual stock transaction amount (*Indaily_turnover*), and turnover rate, respectively.

Table 3: Results of the Basic Regression Test

	(1) <i>chg</i>	(2) <i>Indaily_volume</i>	(3) <i>Indaily_turnover</i>	(4) <i>turnover_rate</i>
<i>lnSVI</i>	0.468*** (16.86)	1.271*** (38.09)	1.287*** (38.42)	5.039*** (27.86)
<i>lnmarket_value</i>	0.949*** (14.76)	0.642*** (10.74)	1.595*** (26.36)	1.887*** (8.29)
<i>mk_chg</i>	0.825*** (108.55)	0.0141*** (13.15)	0.0115*** (10.64)	-0.0395*** (-8.37)
<i>_cons</i>	-24.95*** (-17.87)	-7.355*** (-5.86)	-27.14*** (-21.35)	-75.50*** (-15.18)
<i>Individual-effect</i>	Control	Control	Control	Control
<i>Time-effect</i>	Control	Control	Control	Control
<i>F</i>	4568.53	1647.98	2440.05	383.55
<i>N</i>	465020	465020	465020	465020
<i>Adj R2</i>	0.0176	0.4303	0.5705	0.0014

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

As shown in Table 3, after controlling for individual and time effects, this paper finds that the internet search indexes that substitute for investor sentiment are all significant and positively related to each of the explanatory variables at the 1% level, implying that the internet search indexes (*lnSVI*) all positively affect individual stock price fluctuation (*chg*), individual stock volume (*Indaily_volume*), individual stock transaction amount (*Indaily_turnover*), and turnover rate. Overall, every 1 standard deviation increase in the internet search index increases stock returns by about 0.5 percentage points, with the highest significance between the internet search index (*lnSVI*) and individual stock volume (*Indaily_volume*) and individual stock transaction amount (*Indaily_turnover*), suggesting that the degree of investor's internet attention has a more noticeable impact on individual stock volume and transaction amount. The empirical results in Table 3 corroborate the impact of the internet search index (*lnSVI*) on individual stocks and shares from multiple dimensions, in that the higher the degree of attention paid to individual stocks, that is to say, the higher the investor sentiment,

the higher the stock's return on the same day, and the individual stock volume and turnover are enlarged, and the turnover rate is elevated.

As for the control variables, the total market capitalization of individual stocks (*lnmarket_value*) and stock performance as a whole are both significant at the 1% level, which means that the larger the total market capitalization of individual stocks, the higher the stock return, and the higher the volume, transaction amount, and turnover rate of individual stocks. The rise and fall of the market (*mk_chg*) in which the stock is located also has a significant impact on stock performance. In particular, the significance of market ups and downs (*mk_chg*) versus individual stock ups and downs (*chg*) reaches 108, indicating that market ups and downs (*mk_chg*) are in fact the most important factors affecting individual stock ups and downs (*chg*). Furthermore, there is also a significant positive relationship between market (*mk_chg*) rise and fall and the volume and turnover of individual stocks.

Due to the special characteristics of the stocks of ST companies, investors engaging in speculation on ST companies often have speculative purposes. In addition, financial companies are also not comparable due to the special characteristics of financial statements, so in the next empirical regression we exclude financial and ST companies, leaving 427,753 samples for the study. Similarly, we use a panel model for the regression test and control for individual-effect and time-effect. The regression results are shown in Table 4.

Table 4: Results of the Basic Regression Test (financial industry and ST companies were excluded)

	(1) <i>chg</i>	(1) <i>lndaily_volume</i>	(1) <i>lndaily_turnover</i>	(1) <i>turnover_rate</i>
<i>lnSVI</i>	0.495*** (18.27)	1.304*** (48.60)	1.318*** (49.40)	5.232*** (31.03)
<i>lnmk_value</i>	0.887*** (13.82)	0.551*** (9.72)	1.511*** (26.58)	1.631*** (7.53)
<i>mk_chg</i>	0.823*** (107.41)	0.0133*** (12.19)	0.0107*** (9.75)	-0.0435*** (-9.02)
<i>_cons</i>	-23.63*** (-16.88)	-5.448*** (-4.48)	-25.28*** (-20.68)	-70.53*** (-14.67)
<i>Individual-effect</i>	Control	Control	Control	Control
<i>Time-effect</i>	Control	Control	Control	Control
<i>F</i>	4492.51	2039.75	2865.20	420.33
<i>N</i>	427753	427753	427753	427753
<i>Adj R2</i>	0.0206	0.4246	0.5734	0.0043

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

As shown in Table 4, the results of Table 4 indicate that the regression results of the sample remain similar to the regression results of the full sample after excluding financial and ST companies. The search index (*lnSVI*) remains significantly positively correlated with stock performance at the 1% level, i.e., the search index significantly and positively affects stock price fluctuation (*chg*), stock volume (*lndaily_volume*), stock transaction amount (*lndaily_turnover*), and stock turnover. The results in Table 4 demonstrate the robustness of the empirical model.

5. Conclusion

This paper empirically tests the empirical relationship between investor sentiment and stock returns by utilizing the Internet search index as a proxy variable for investor sentiment. A panel model regression test on data from a sample interval of a total of 242 stock trading days in China's A-share mainboard market in 2023 reveals that the internet search index has a significantly close relationship

with stock market performance. Specifically, internet search index (lnSVI) significantly affects the individual stock price fluctuation (chg), individual stock volume (lndaily_volume), individual stock transaction amount (lndaily_turnover), and turnover rate at the 1% level. This finding provides thorough evidence that investors' emotional state indeed influences their daily investment decision-making behavior: the higher level of investors' emotional concern is often accompanied by a corresponding increase in daily stock returns. This finding deepens our understanding of the complex relationship between investor sentiment and stock market performance and provides valuable reference information for market participants.

China's A-share market presents unique challenges driven by sentiment, policy, and media influence. By staying rational, fostering financial literacy, and using technology judiciously, investors can navigate these complexities effectively. Remember, the goal is to base your investments on facts and sound analysis, not on fleeting emotions or trends. Media sensationalism and spikes in online search activity often indicate emotional responses rather than a rational analysis which could be apparently portrayed in internet search index (lnSVI). Treat online search trends as one of many indicators, not the sole basis for investment decisions. Instead of reacting to headlines, dig deeper into company performance and industry developments. Herd behavior, characterized by “chasing highs and selling lows,” is prevalent among retail investors. Following the crowd often leads to poor outcomes and heightened volatility. Form your own opinions by analyzing credible sources of information. Developing a long-term perspective to resist short-term market hype is the key point when investing.

References

- [1] Stein, J. C. (1996). *Rational capital budgeting in an irrational world*. *Journal of Business*, 69(4), 429-451.
- [2] Brown, G. W., & Cliff, M. T. (2005). *Investor sentiment and the near-term stock market*. *Journal of Empirical Finance*, 12(1), 1-28.
- [3] Wurgler, J., & Baker, M. (2006). *Investor sentiment and the cross-section of stock returns*. *Journal of Finance*, 61(5), 1645-1680.
- [4] Wen, Y., Zhu, X., & Chen, L. (2014). *The impact of investor sentiment on stock market returns: Evidence from China*. *Journal of Financial Research*, 37(3), 42-53.
- [5] Wang, C., & Wang, J. M. (2020). *Investor attention and its impact on stock returns and trading volume*. *Journal of Financial Studies*, 45(2), 24-39.
- [6] Da, Z., Engelberg, J., & Gao, P. (2011). *In search of attention*. *Journal of Finance*, 66(5), 1461-1499.
- [7] Liu, Y. (2018). *Investor attention and stock market: Evidence from Baidu Index, 360 Index, and He Xun Index*. *Economic Research Journal*, 53(2), 78-92.
- [8] Khan, M. M., Sadiq, M., & Wang, J. (2020). *Investor sentiment and stock market: Using Google Trends data to construct sentiment indices*. *International Journal of Finance & Economics*, 25(4), 819-834.