

Investor Sentiment and Tech-Sector Stock Returns: An Empirical Analysis Based on Guba Comment Volume

Junyi He^{1a*},
Wangzheng Wanyan^{1b},
Huandong Li^{1c},
Chenkai Zhu^{1d}

^{1a*} College of Management and Economics, Tianjin University, Tianjin, 300072, China, 1940605761@qq.com

^{1b} School of Economics and Management, China University of Geosciences (Wuhan), Wuhan, 430074, China, 15614668856@163.com

^{1c} College of Business and Public Management, Wenzhou-Kean University, Wenzhou, 325060, China, 1517625083@qq.com

^{1d} Santiago Canyon College, Orange County, 92869, U.S., 2703757827@qq.com

*Corresponding author email: 1940605761@qq.com

Abstract:

Purpose - This study investigates whether investor sentiment carries causal information about the future returns of five large-cap Chinese technology stocks.

Design/methodology/approach - High-frequency proxies for sentiment are built from the daily count of emotion-tagged posts on Guba and from daily return metrics. After cross-sectional correlation and ADF stationarity checks, two-stage residual regressions and pooled panel specifications with lagged sentiment terms are estimated. The sample is then split into training and validation windows; out-of-sample forecasts are benchmarked against the Efficient-Market-Hypothesis (EMH) null. Finally, mean-variance frontiers that blend the five equities with Chinese government bonds are constructed to obtain implementable portfolio weights.

Findings - Non-contemporaneous sentiment exerts a statistically and economically significant effect on next-day returns. Both the residual and pooled models deliver out-of-sample MSPEs that are materially lower than the EMH benchmark. Incorporating sentiment-adjusted expected returns shifts the efficient frontier north-west, yielding higher Sharpe ratios.

Originality - The paper is the first to fuse intraday social-media emotion data with conventional econometric tests of predictability for Chinese technology firms, and to show that sentiment-driven return forecasts can be translated into realizable portfolio allocations.

Keywords: - investor sentiment; stock returns; portfolio optimization; time-series analysis

1. Introduction

1.1 Research Background

The ongoing development of financial markets means that stock market volatility is influenced by numerous factors, with investor sentiment playing a crucial role. A broad empirical literature establishes that shifts in investor sentiment are closely linked to time-varying patterns of stock return volatility. For example, during the onset of the COVID-19 pandemic in early 2020, major global equity indices exhibited extreme volatility, with the S&P 500 recording a 34% decline within a single month and the VIX surging to its highest level since the 2008 financial crisis [1]. Within a single month, the S&P 500 plunged roughly 34 per cent and the VIX volatility index surged from approximately 12 to 82—its highest reading in years—underscoring extreme investor fear [2]. This evidence clearly shows that extreme investor panic exerts significant pressure on stock markets [3].

Traditional finance theory often emphasises market efficiency, largely neglecting the influence of investor psychology on asset pricing. However, recent advancements in big data and time-series analysis have made the incorporation of investor sentiment indicators into stock return models increasingly crucial in financial research. Stock returns are affected by various factors, with investor sentiment playing a key role. This is evidenced by the correlation between sentiment swings and market movements, as well as their impact on individual stocks. The 2016 U.S. presidential election provides a salient illustration: shifts in investor sentiment, triggered by policy-related expectations and confidence swings surrounding the candidates, materially contributed to equity-market volatility [4]. Similarly, sentiment can impact stocks in specific sectors or companies, such as the environmental sector, where shifts in ecological policies may affect investor outlook and, in turn, returns. In recent years, both academic and professional circles have increasingly focused on the importance of investor sentiment. Data indicate that short-term stock returns are clearly affected by shifts in sentiment [5].

1.2 Research Purpose

This study aims to employ time series methodologies to comprehensively investigate the underlying mechanisms through which investor sentiment influences stock returns across different time horizons [6]. By characterising the dynamic nexus between sentiment fluctuations and subsequent returns, the study furnishes investors and regulators with rigorous, evidence-based benchmarks for decision-making. This contributes to enhancing market stability and investment efficiency [7]. Theoretically, the paper extends the analytical framework linking investor

sentiment to equity returns, offering novel insights that advance the frontier of financial research. For example, historical data analysis reveals that specific shifts in investor sentiment often precede pronounced fluctuations in stock market returns, empirically confirming the influence of sentiment on the market.

1.3 Research Significance

1.3.1 Emotion Measurement Innovation

This paper utilises the daily counts of positive, neutral, and negative comments for each stock from GUBA in the CNRDS database to construct three distinct indicators for quantifying investor sentiment. This represents an innovation relative to existing studies, most of which infer investor sentiment from publicly disclosed financial statement data or trading statistics provided by stock software. Such approaches are outcome-oriented and overlook the subjective nature of investor emotions. In contrast, this paper directly analyses investor comments. By constructing three sentiment proxies within distinct regression frameworks, the study embeds investors' subjective affective states and broadens the model's admissible parameter space, yielding more robust evidence that sentiment significantly influences equity returns.

1.3.2 Test for Asymmetric Sentiment Effects

By constructing and empirically examining three separate metrics—positive sentiment, negative sentiment, and overall sentiment—this study sensitively detects whether the effects of investors' positive versus negative emotions on stock returns are symmetric and which emotion exerts a more pronounced impact. The empirical results reveal that investors' negative sentiment exerts a larger and significantly negative influence on stock returns; this finding offers firms actionable guidance for maintaining share prices and enhancing equity value—namely, minimising the proportion of investors' negative sentiment.

1.3.3 Forecasting the Superiority of Models

The study partitions the sample into a training set and a test set, estimates the regression equation on the former, and then forecasts stock returns for the latter. Prediction errors are computed by comparing the predicted values with the realised returns. Relative to forecasts generated under the Efficient Market Hypothesis (EMH), the proposed model yields minor prediction errors and higher accuracy. Such a model not only furnishes investors with a more reliable stock-selection criterion but also supplies listed firms with operational guidance for managing return volatility.

1.3.4 Research Implementation

Grounded in Markowitz's mean-variance framework and using Chinese government bond yields as the risk-free

rate, this paper constructs optimal portfolios by combining the return series of five representative technology firms. The procedure yields two allocation schemes—maximum Sharpe ratio and minimum variance—thereby translating the empirical findings into implementable strategies. The results provide stock exchanges with a sentiment factor, facilitating the design of “retail-sentiment-neutral” or “contrarian-sentiment” ETFs.

1.4 Research Questions

This study empirically examines the extent to which contemporaneous and lagged investor-sentiment shocks—quantified via the sentiment index constructed from Guba comments—explain the time-varying excess returns of Chinese technology stocks, and evaluates whether, after controlling for standard risk factors and market microstructure effects, sentiment-augmented time-series specifications outperform the baseline Efficient-Market-Hypothesis (EMH) model in out-of-sample forecasting accuracy, as measured by the out-of-sample R^2 and the Clark–West statistic.

2. Literature Review

2.1 Emotion and Financial Decision-Making

Cognitive psychology demonstrates that decision-making is not solely derived from rational deliberation; it is also shaped by affective states, perceptions, and attitudes [8]. Loewenstein et al. (2001) [9] observe that, in actual decisions, emotional reactions under risk frequently diverge from cognitive appraisals, and when such divergence arises, affect often dominates behavioural outcomes. Because investment choices are inherently risky and uncertain, emotions likewise influence investors’ decisions [10]. Within behavioural finance, sentiment is regarded as an essential endogenous driver of asset-price fluctuations [11]; theoretically, transient deviations of prices from fundamental values are attributed to investors’ irrational trades and affective swings, which serve as endogenous sources of short-run market volatility [12]. LAM et al. (2010) [13] quantify investors’ risk preferences via the representativeness and conservatism heuristics, and, employing a quasi-Bayesian learning framework, demonstrate that the representativeness heuristic overweights recent clusters of observations and underweights older observations. In contrast, the conservatism heuristic underweights recent information and overweights prior information.

Taken together, the existing literature on how emotion affects financial decision-making primarily focuses on how sentiment influences retail investors’ choices in market settings, relying predominantly on large, long-horizon samples for empirical analyses. Systematic investiga-

tion of short-horizon regularities remains comparatively scarce.

2.2 Sentiment and Stock Returns

A broad scholarly consensus holds that sentiment and stock returns are linked. Empirical inquiries into this nexus can be grouped into two strands. The first constructs investor sentiment proxies from publicly available, official firm-level indicators and then examines their statistical relationship with returns. Baker & Wurgler (2006) [14] create a composite sentiment index by extracting the first principal component from six market-based proxies: the closed-end fund discount, NYSE turnover, IPO volume and average first-day returns, the equity share of new issues, and the dividend premium. Their study shows that investor sentiment exerts a statistically significant effect on the cross-section of subsequent stock returns, with effects that vary systematically across firms with heterogeneous characteristics. When initial sentiment is elevated, stocks that appeal to optimists and speculators but are challenging to arbitrage subsequently earn lower returns. Fuwei Jiang, Xiumin Martin, Joshua Lee, and Guofu Zhou (2019) [15] construct managerial sentiment indicators from the tone of conference calls and financial statements; employing standard predictive regressions, they document that managerial sentiment negatively forecasts aggregate earnings growth—high managerial sentiment is associated with lower aggregate growth in the following year. Kahneman & Tversky (2013) [16] argue that the marginal disutility of losses exceeds the utility of equivalent gains, implying that investors trade more aggressively in response to adverse information. Shleifer & Vishny (1997) [17] demonstrate that capital constraints become tighter after price declines, resulting in negative sentiment that is less fully offset by arbitrage activity. Paul C. Tetlock (2007) [18] estimates a VAR model and finds that a pessimism factor exerts a significant negative contemporaneous impact on the Dow Jones Industrial Average; however, this effect is temporary and is almost entirely reversed later in the trading week.

The second strand leverages textual analysis, classifying the text of investor posts or corporate annual reports into positive and negative categories to construct sentiment measures, or directly training machine-learning models that map sentiment-laden text to stock returns. In Zhang et al. (2017) [19], a corpus of investor comments is created, and a bidirectional LSTM model is used to classify the sentiment embedded in these comments. The trained model achieves high out-of-sample accuracy in predicting stock prices. Silva de Carvalho et al. (2023)[20] observe, in their systematic literature review, that most studies treating textual content as a direct sentiment measure employ dictionary-based or semantic fingerprint approaches. The coexistence of these two methodological paradigms

attests to the richness and maturity of extant operationalisations of sentiment; accordingly, the present study will synthesise the respective strengths of both approaches in constructing its sentiment indicators.

We searched the Web of Science and retrieved 2,833 relevant articles. After exporting the keyword information, we performed a co-occurrence analysis with VOS viewer; the resulting map (Fig. 1) shows that in the sentiment-centred

literature, the terms “news” and “media” appear, confirming that prior studies have used news and media information to gauge investor sentiment. The simultaneous appearance of “consumer confidence,” “expectations,” and “overconfidence” indicates that scholars have investigated whether markets characterised by pronounced sentiment can generate abnormal returns.

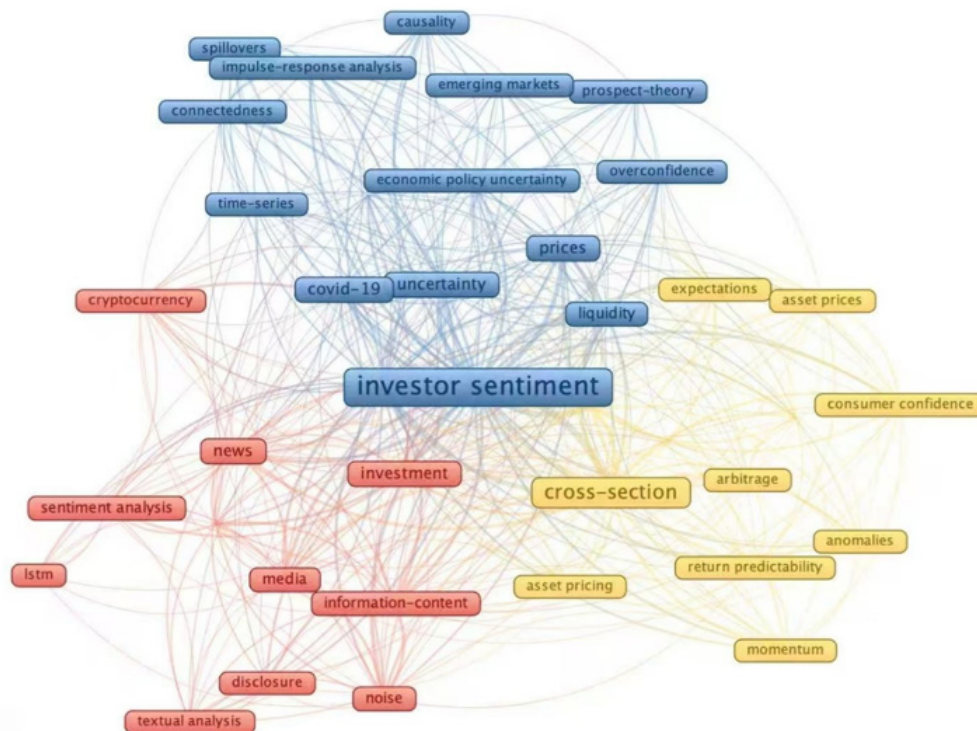


Figure 1. Co-occurrence network of keywords

Collectively, the extant literature predominantly examines the contemporaneous impact of sentiment on returns, regressing current sentiment measures on current returns. Investigations of how sentiment in other periods influences current returns—i.e., the dynamic effect of sentiment on returns—remain scarce.

2.3 Sentiment and Equity-Return Forecasting

Classical finance theory postulates investor rationality, yet empirical reality reveals that trading decisions are systematically impeded by a constellation of behavioural and psychological frictions [8]. Because investor sentiment constitutes a salient driver of non-rational trading activity and has been documented to influence stock returns, its dynamic relation with returns can be exploited to forecast short-horizon return fluctuations. Early work by Lai et al. (2014) [21] surveys a broad body of empirical studies. It

concludes that both subjective survey-based and objective market-based sentiment proxies exhibit predictive content for near-term returns. Examining Korean KOSPI data, Kim and Ryu (2020) [22] establish that investor sentiment delivers superior out-of-sample predictive power for equity returns relative to a comprehensive set of macroeconomic indicators; the term spread, a counter-cyclical macroeconomic indicator, predicts stock returns only at long horizons. The methodological taxonomy for equity-return forecasting mirrors the dichotomy described in Section 2.2: one strand relies on sentiment quantification, whereas the other adopts machine-learning techniques; a nascent literature further integrates both approaches. Existing studies continue to emphasise the predictive power of contemporaneous sentiment measures, whereas the forecasting efficacy of non-concurrent sentiment has received limited attention. Additionally, extant sentiment-quantification studies predominantly map sentiment proxies

to returns via benchmark factor models (e.g., CAPM or Fama–French), so the robustness of predictive power to market-wide risk adjustments remains untested.

3. Data and Research Design

3.1 Data Source

We get the data from the Chinese Research Data Services (CNRDS) Platform. We retrieved the full population of Guba comments directed at A-share listed firms from August 2, 2023, to September 30, 2024, including daily counts of positive, negative, and total posts. Daily stock returns and the CSI 300 index were concurrently downloaded from Flush Financial Terminal to facilitate the subsequent empirical analyses.

3.2 Data Cleaning

Given that stock returns are recorded exclusively on trading days. In contrast, post volumes are tallied on both trading and non-trading days; we aligned the datasets in Python by treating the return series as the master time index. We performed an inner merge, thereby retaining only observations corresponding to official trading sessions.

3.3 Sample Selection

We select BYD, Hikvision, iFLYTEK, Contemporary Amperex Technology Co. (CATL), and Semiconductor Manufacturing International Corporation (SMIC) as the sample firms. The selection is based on two considerations. First, these five companies operate in key segments of China's technology market—new energy vehicles, video surveillance, artificial intelligence and natural-language processing, power batteries, and domestic semiconductors—that receive intense market attention and play a leading role in the nation's technological development. BYD is the global top-selling new-energy-vehicle manufacturer; Hikvision commands the leading international market share in the design and delivery of video-surveillance solutions; iFLYTEK ranks first in China for AI natural-language processing; CATL records the highest global installed capacity of power batteries; and SMIC is the most technologically advanced wafer foundry in mainland China. As obvious competitors, all five are among the most discussed stocks on the Guba forum, yielding comprehensive data. Their return dynamics therefore reflect not only firm-specific fundamentals but also market sentiment toward China's

major technology sectors and policy expectations, providing a valuable laboratory for research. Second, the investor base of each firm is primarily comprised of retail traders, and no single shareholder holds more than 50% of the outstanding shares. Retail trades are more likely to be driven by sentiment, and these sentiment effects are not neutralised by a controlling shareholder, so that they can be detected in price and return movements. Consequently, the ownership structure of the five companies is well-suited to the objectives of this study.

The empirical window spans from August 2, 2023, to September 30, 2024. The period from August 2, 2023, to May 20, 2024, is used for estimation and model training, while the period from May 21 to September 30, 2024, is reserved for out-of-sample testing. As argued earlier, sentiment exerts a short-lived shock on returns. Because agents gradually forget past emotions, the return distortion attributable to sentiment diminishes as the horizon lengthens, making it challenging to identify. Restricting the estimation window to a short, recent interval mitigates the decay of sentiment-induced mispricing, thereby enhancing the identifiability of retail sentiment's effect on subsequent returns.

3.4 Variables Definition and Calculation

3.4.1 Measuring Investor Sentiment

We constructed the Stock Sentiment Leaning Index (SSLI) to measure investor sentiment. The index is computed according to the following expression:

where SSLI is the measurement of investor sentiment. The formula indicates that when SSLI is less than 0, it may reflect a negative attitude towards the daily returns. Additionally, when SSLI is greater than 0, it indicates a positive attitude toward daily returns.

Analogously, we constructed unilateral sentiment proxies by separately scaling the daily counts of positive and negative posts:

where PO and NE are the percentages of positive posts and negative posts, respectively.

3.4.2 Measuring Stock Returns

Daily returns serve as the benchmark for gauging individual stock performance,

where r_t means the return rate of the stock at time t , P_t means the closing price of the stock at time t .

In our study, we also used the following variables, as shown in Table 1.

Table 1 Variable Definition

Variables	Symbol	Definition
Individual stock return		Stock i's daily return
Abnormal return		Abnormal return after market adjustment
Market return		Market index daily return
SSLI		Stock Sentiment leaning index
PO		Percentage of positive posts
NE		Percentage of negative posts

3.5 Regression Models Construction

To explore the relationships between the variables of interest in our research, we employ three distinct models. Each model is designed to capture different aspects of the dynamic sentiment's influence on stock returns, enabling a comprehensive analysis. The following presents these models in detail.

3.5.1 Two-Stage Residual Model I

To identify the dynamic effect of investor attention on firm-specific returns, we implement a Two-Stage Residual Model (TSRM). In the first stage, we purge everyday market movements by estimating a daily single-factor market model for each stock :

where r_{it} means the return of stock i on day t and r_{mt} is the contemporaneous market return. The residual ϵ_{it} captures the abnormal returns of stocks.

In the second stage, we regress these residuals on lagged values of an attention proxy in an autoregressive-distributed-lag framework:

where $\bar{\epsilon}_{it}$ is sentiment for stock i on day t , $\bar{\epsilon}_{it}$ is the sample mean for stock i , and μ is an error term. We also used the lagged term of abnormal return (ϵ_{it-l}) to determine whether there was a lag effect. The demeaning of ϵ_{it} is performed to eliminate stock-fixed effects and to isolate the idiosyncratic component of sentiment shocks. The coefficients α trace the impulse-response of returns to attention shocks up to three trading days after the shock, while β controls for return autocorrelation.

3.5.2 Two-Stage Residual Model II

To disentangle the effects of positive versus negative sentiment, we extend the baseline framework to Two-Stage Residual Model II (TSRM-II). In Stage 1, the market model is again estimated:

Stage 2 then substitutes the idiosyncratic residuals into two separate specifications:

where ϵ_{it}^+ denotes the proportion of positive (negative) posts for stock i on day t , and $\bar{\epsilon}_{it}^+$ is its sample mean. Demeaning isolates the unexpected component of sentiment, with the three lags capturing the short-term propagation and potential reversal of sentiment effects. Estimation

follows the same rolling-window and Newey–West error structure described for TSRM-I.

3.5.3 Pooled Regression

To provide a parsimonious benchmark, we also estimate a pooled panel regression:

where ϵ_{it} represents the composite sentiment index (SSLI). The term ϵ_{it} controls for first-order return autocorrelation, and two-way clustered standard errors (firm and time) are employed to account for cross-sectional and serial dependence.

3.6 Descriptive Statistics

3.6.1 Descriptive Statistics for Core Variables

Table 2 tabulates the descriptive statistics for the key variables. According to Panel A, the mean and median of the SSLI are closely aligned, indicating an approximately symmetric distribution across the sample; the mode converges to 0, suggesting that short-horizon assessments are concentrated around neutrality; cross-sectional standard deviations differ markedly: BYD (id = 1), iFLYTEK (id = 3), and SMIC (id = 5) exhibit relatively low sentiment volatility, whereas Hikvision (id = 2) records the highest; the range between the maximum and minimum observations is widest for Hikvision, including extremely negative sentiment; the other 4 stocks display narrower intervals with smaller absolute extremes, implying that negative sentiment is more prone to tail events and may exert a disproportionately larger effect on subsequent returns. Turning to Panel B, average daily returns are statistically hovering narrowly around 0, consistent with a short-horizon random walk, which facilitates the identification of sentiment effects; median returns are likewise close to 0, corroborating distributional symmetry; SMIC (id = 5) exhibits the most significant return volatility, whereas Hikvision (id = 2), despite displaying the largest sentiment variation, records the lowest, underscoring that sentiment amplitude and return variance do not move in lockstep; comparing the absolute magnitudes of the extremes, the largest positive realization exceeds the absolute minimum, providing preliminary evidence of momentum-chasing

behaviour among investors.

Table 2 Descriptive Statistic of SSLI and Returns

Variable	N	Mean	Median	Mode	Std	Min	Max
Panel A: Descriptive statistics of SSLI for each stock							
	426	0.304	0.333	0.333	0.241	-0.371	0.778
	423	0.057	0.042	0	0.311	-1	0.846
	426	0.224	0.236	0.2	0.212	-0.427	0.722
	426	0.184	0.221	0	0.291	-0.460	0.768
	426	0.206	0.222	0	0.267	-0.565	0.895
Panel B: Descriptive statistics of Return Rate for each stock							
	283	0.000	-0.001	-0.010	0.020	-0.066	0.084
	283	-0.000	-0.000	0.000	0.017	-0.069	0.088
	283	-0.000	-0.002	-0.023	0.028	-0.100	0.100
	283	0.001	-0.002	-0.011	0.026	-0.147	0.187
	283	0.002	-0.000	0.004	0.032	-0.106	0.2

3.6.2 Data Distribution

Fig. 2-Fig. 6 present line graphs showing the daily trajectories of the Stock-Specific Sentiment Level Index (SSLI) and the related returns for each of the five sample firms. Across all five stocks, both the SSLI and return series show no clear long-term trend, with short-term fluctuations dominating. This pattern indicates that firm-specific developments and temporary market events tend to have a more immediate effect on returns. Towards the end of September, all five stocks experienced notable upward

price movements, a rise we tentatively link to the near-simultaneous policy easing by the Federal Reserve and the People's Bank of China in late September 2024. Since this event is brief but macroeconomically significant, with direct implications for asset pricing, we decide to keep the observations rather than dismissing them as outliers. Throughout most of the sample, SSLI and return volatility move together—calm sentiment corresponds with modest returns, while turbulent sentiment aligns with larger price swings—indicating a contemporaneous connection between sentiment and returns.

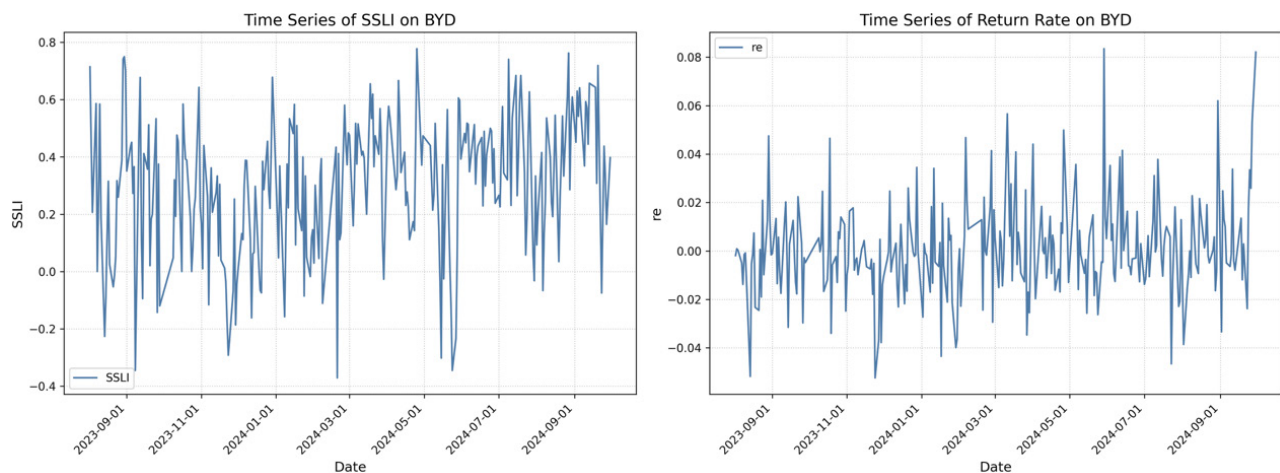


Figure 2. Daily SSLI and Return Trends on BYD

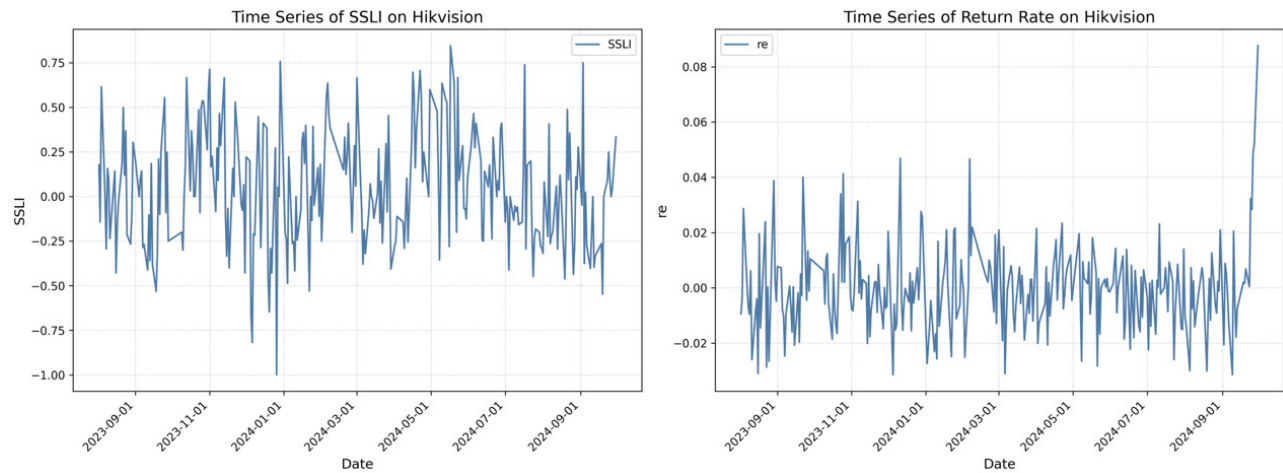


Figure 3. Daily SSLI and Return Trends on Hikvision

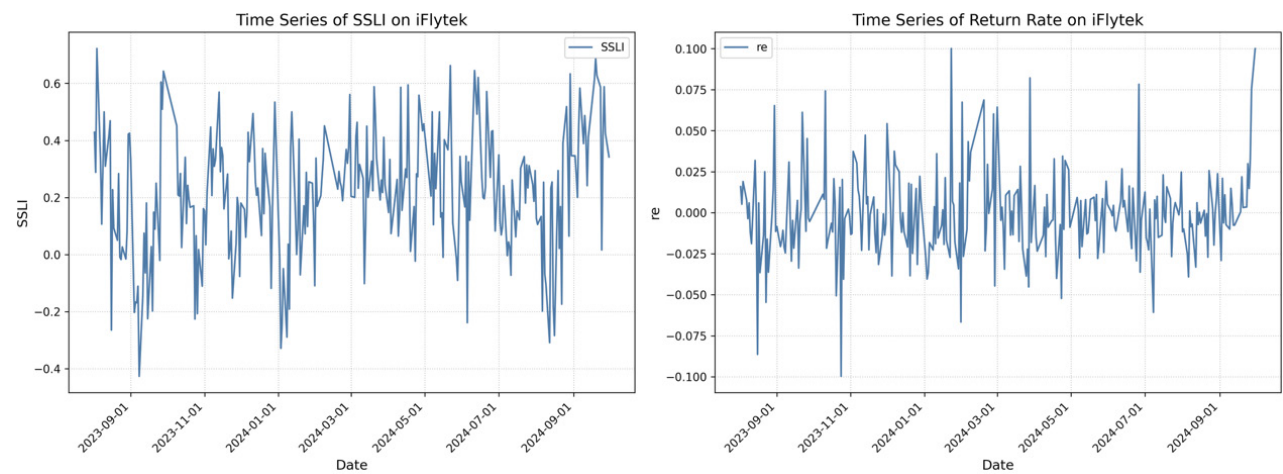


Figure 4. Daily SSLI and Return Trends on iFLYTEK

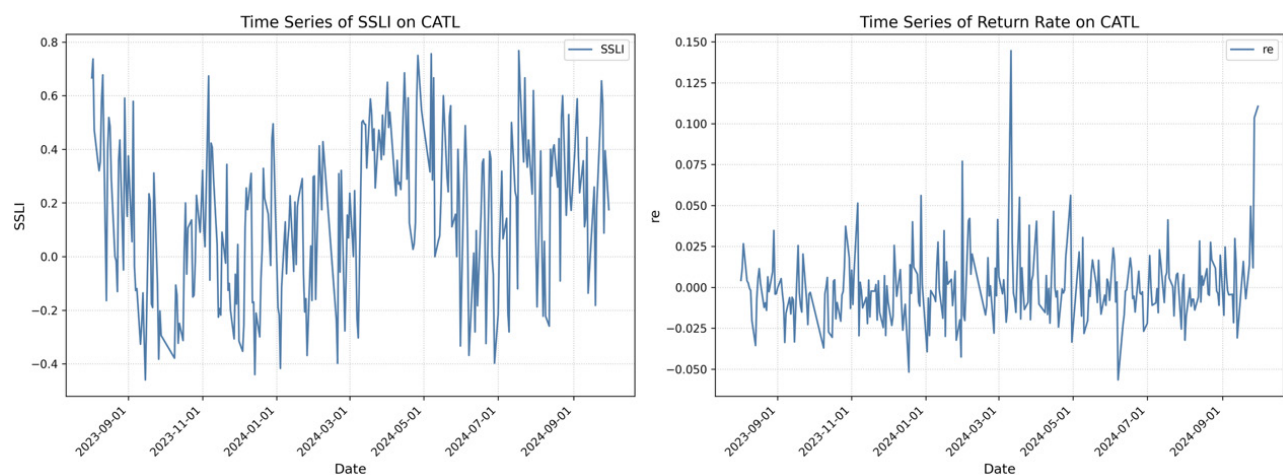


Figure 5. Daily SSLI and Return Trends on CATL

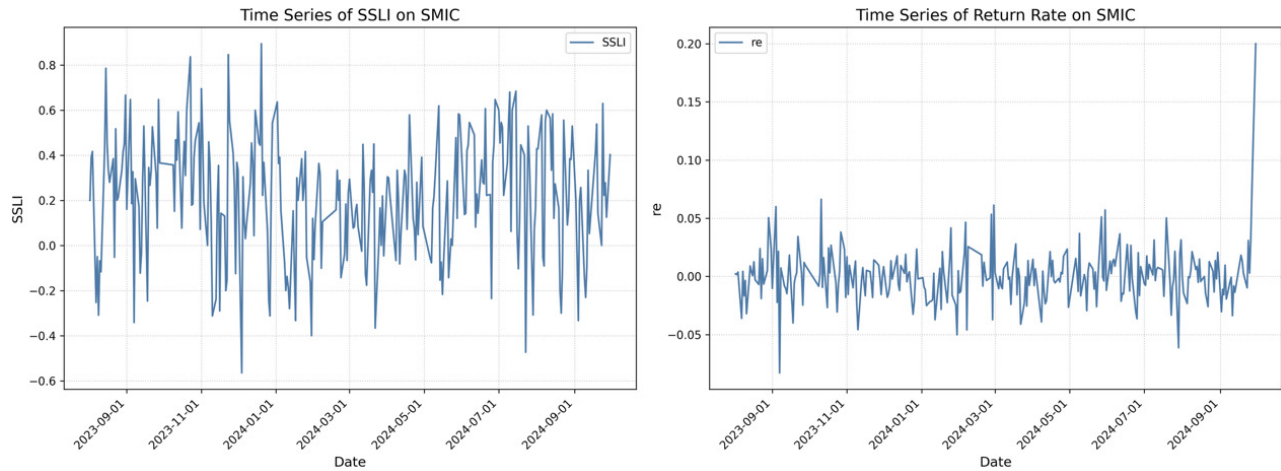


Figure 6. Daily SSLI and Return Trends on SMIC

3.6.3 Stationarity Tests

To prevent spurious regression, we perform Augmented Dickey–Fuller (ADF) unit-root tests on all series. The

results, presented in Table 3, indicate that all p-values are well below 0.05, confirming that each variable is covariance-stationary. Therefore, the risk of spurious regression is effectively ruled out.

Table 3 Results of ADF Test

Variable	Test Statistic	P-value
Panel A: ADF test results of SSLI		
	-10.977	0.000
	-12.088	0.000
	-10.283	0.000
	-8.964	0.000
	-12.476	0.000
Panel B: ADF test results of return rate		
	-17.650	0.000
	-16.383	0.000
	-19.463	0.000
	-22.686	0.000
	-25.198	0.000

3.6.4 Correlation Matrix

A correlation matrix—also known as a cross-correlation matrix—is created by calculating the pairwise Pearson correlation coefficients between all variables. A heat map offers an intuitive visualisation: numerical correlations are represented by a colour gradient, with darker hues indicating stronger connections between stocks. Macroeconomic events and industry-level policies common to the equity market cause co-movements in stock indicators within the same sector, thus increasing short-term cross-industry differences. Therefore, correlation analysis helps to filter out shared factors and tune models. At the same time, overlaps, competition, and cooperation among firms pro-

duce inter-stock influences, allowing correlation analysis to identify valid connections.

Fig. 7 presents the heat map of SSLI and returns for the five sample stocks. The left-hand heat map indicates that inter-stock sentiment correlations are modest, ranging mainly between -0.20 and 0.40 . SMIC exhibits the lowest correlations with the remaining firms, whereas CATL and BYD register the highest pairwise coefficient of 0.32 ; as leading competitors in the new-energy segment, both companies attract intensive retail attention, generating synchronised sentiment dynamics. The right-hand heat map shows that return correlations are markedly higher, clustering between 0.30 and 0.70 . BYD and CATL display the strongest co-movement at 0.64 . Taken together

with their SSLI correlation of 0.32, the evidence suggests a potential transmission mechanism whereby sentiment

co-movement influences return co-movement.

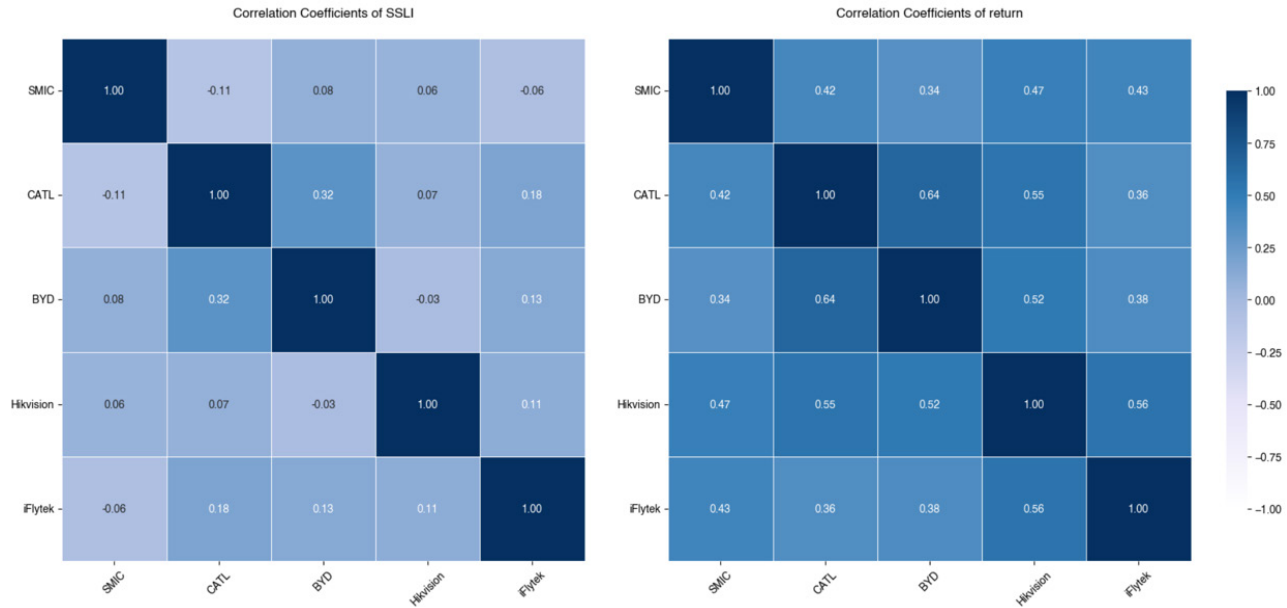


Figure 7. Heat Maps of Correlation Coefficients between SSLI and Returns

4. Empirical Analysis

4.1 Empirical Results

4.1.1 Two-Stage Residual Model I

Table 4 presents the regression results of the Two-Stage Residual Model I. After estimating the market model in the first stage, the risk-adjusted returns (i.e., the residuals from the first-stage regression) are used in the second stage. The second-stage results show that the one-day-lagged SSLI deviation has a positive but insignificant effect on risk-adjusted returns at the 5 per cent level; the two-day-lagged SSLI deviation has a positive and significant impact at the 5 per cent level; and the three-day-

lagged SSLI deviation has a negative and significant effect at the 5 per cent level. The results indicate that sentiment's influence on stock returns reverses after three trading days. We conjecture that this reversal occurs because, on the event day and the second day, trading is dominated by the sentiment of retail investors, whose net effect continues to influence returns. In contrast, on the third day, earlier momentum buyers begin to take profits and part of the sentiment deviation is perceived as an error, creating an opposing shock that turns the coefficient negative. Additionally, since the SSLI captures the net effect of positive sentiment on returns, the reversal on the third day may reflect that the positive sentiment's influence is starting to be outweighed by negative sentiment, a possibility confirmed in Two-Stage Residual Model II.

Table 4 Regression Results of Two-Stage Model I

VARIABLES						
	1.380*** (0.000)	1.072*** (0.000)	1.479*** (0.000)	1.575*** (0.000)	1.306*** (0.000)	
						-0.113* (0.092)
						0.007 (0.129)
						0.005** (0.031)
						-0.010** (0.014)

Constant	-0.000 (0.784)	0.000 (0.813)	-0.001 (0.669)	0.000 (0.948)	-0.000 (0.869)	-0.001*** (0.000)
Observations	191	191	191	191	191	355
R-squared	0.445	0.406	0.220	0.383	0.318	0.042

The table presents the estimation results for the Two-Stage Residual Model I. r_{it} denotes the return of stock i . ϵ_{it} is the residual obtained from the first-stage regression, representing the risk-adjusted excess return of stock i . r_{mt} denotes the market return, proxied by the CSI 300 Index. The SSLI (Stock Sentiment Leaning Index) is a proxy for retail investor sentiment. In each cell, the first figure presents the estimated coefficient, and the second displays the corresponding p-value. Statistical significance levels are indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.1.2 Two-Stage Residual Model II

Because Stage 1 of the two-stage residual Model 2 is identical to that of Model 1, the relevant estimation results are omitted here for brevity. Table 5 shows the regression results of the second stage. In the specification with the positive-post ratio as the sentiment proxy, only the three-period-lagged deviation in this ratio has a margin-

ally significant effect on risk-adjusted returns; all other variables fail to reach conventional significance levels. In the specification employing the negative-post ratio, deviations at lags two and three are negatively and significantly related to subsequent risk-adjusted returns ($p < 0.05$). Notably, the lagged risk-adjusted return itself is insignificant in both models, implying that return persistence is absent once sentiment effects are controlled for. Compared to its positive-sentiment counterpart, the negative-post-ratio specification includes more significant regressors and exhibits a higher R-squared value, indicating that negative sentiment has a stronger economic impact on stock returns than positive sentiment, a pattern consistent with the findings reported in Two-Stage Residual Model I. Consequently, all subsequent forecasting exercises use the model that employs the negative-posting ratio as the sentiment proxy.

Table 5 Regression Results of Two-Stage Model II

Variables	(1)	(2)
	-0.103 (0.119)	-0.101 (0.126)
	0.025 (0.233)	
	0.008 (0.493)	
	-0.018* (0.051)	
		-0.025* (0.063)
		-0.030*** (0.000)
		0.042** (0.028)
Constant	-0.001*** (0.000)	-0.001*** (0.000)
Observations	355	355
R-squared	0.024	0.048

Table provides the second-stage regression results of the Two-Stage Residual Model II. α denotes the proportion of positive comments for each stock, whereas β denotes the proportion of negative comments. In each cell, the first

figure provides the estimated coefficient, and the second represents the corresponding p-value. Statistical significance is marked as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.1.3 Pooled Regression

Table 6 presents the pooled regression results. We found that the model was highly significant, indicating that both coefficients and were significantly different from zero. This implies that the model exhibits inertia, as lagged returns significantly influence subsequent-period perfor-

mance. At the same time, the model also has a reversal effect: when the current yield is too high or too low, the yield in the next period will fluctuate in the opposite direction. The results of lag-two sentiments indicate that the positive sentiment index has a positive influence on yield and aligns with the direction of yield change, thereby supporting our hypothesis.

Table 6 Regression Results of Pooled Regression

Variables	
	0.0085*** (0.001)
	-0.1929*** (0.000)
Constant	-0.0005 (0.480)
Observations	750
R-squared	0.028

The table presents the pooled regression estimates. SSLI denotes the Stock Sentiment Leaning Index, and return denotes the stock return. Statistical significance is indicated as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.2 Comparative Stock Return Predictions across the Models above

The prior findings establish that investor sentiment exerts a statistically significant but short-lived effect on equity returns. Consequently, incorporating sentiment-based return forecasts into the investment process can mitigate losses arising from mispricing [23].

The data from August 2, 2023, to May 20, 2024, is selected as the training set, and the data from May 20 to September 30, 2024, is chosen as the test set. Subsequently, we use a recursive method to make predictions, subtracting the actual value of the test set from the predicted value and calculating the sum of the squares of the residuals. Its formula is as follows:

The coefficient of determination (R^2) can measure the precision of the model's predictions. The smaller the score, the higher the prediction accuracy.

Then, we take the well-known efficient market hypothesis (EMH) model as the comparison object, and its prediction formula is as follows:

where $\hat{r}_t$ means the predicted return rate of stock i at time t , r_t means the real return rate of that, $\bar{r}_t$ means the mean return rate of stock i over the periods before the current time.

In the two-stage residual regression, the estimated coefficients are used in the forecasting process: Stage 2 is first projected to produce the predicted risk-adjusted returns (i.e., the Stage-1 residuals), which are then substituted into Stage 1 to recover the final return forecasts. We then calculate (i) daily prediction errors, (ii) the sum of squared errors (SSE) for each stock, and (iii) the overall SSE across all stocks. Forecast accuracy is assessed by comparing the mean absolute error (MAE) and SSE: the smaller the error, the greater the predictive power. In the pooled regression, we directly apply the estimated coefficients to the dataset's lag-one sentiment proxy and lag-one returns to forecast the contemporaneous return.

Table 7 reports the forecast errors for Two-Stage Regression Model I, Two-Stage Regression Model II, the Pooled Regression, and the EMH benchmark. It indicates that both two-stage residual models yield the smallest sum of squared errors, mean squared error, and root mean squared error—whether evaluated stock by stock or in aggregate—demonstrating superior predictive accuracy. This finding supports the conclusion that incorporating sentiment indicators into the forecast of risk-adjusted returns significantly enhances precision. The pooled regression and the EMH benchmark yield almost identical forecast errors, both of which are markedly higher than those of the two-stage residual models. Such a discrepancy indicates that markets deviate from full efficiency and that retail sentiment exerts a systematic, non-negligible effect on equity returns.

Table 7 Results of Prediction of all models

Model		SSE	MSE	RMSE
Two-Stage Model I	BYD	0.0359	0.0004	0.0198
	Hikvision	0.0108	0.0001	0.0108
	iFLYTEK	0.0222	0.0002	0.0155
	CATL	0.0260	0.0003	0.0168
	SMIC	0.0375	0.0004	0.0202
	SUM	0.1324	0.0003	0.0170
Two-Stage Model II	BYD	0.0371	0.0004	0.0201
	Hikvision	0.0108	0.0001	0.0108
	iFLYTEK	0.0223	0.0002	0.0156
	CATL	0.0263	0.0003	0.0169
	SMIC	0.0376	0.0004	0.0202
	SUM	0.1341	0.0003	0.0171
Pooled Regression	BYD	0.0458	0.0005	0.0223
	Hikvision	0.0275	0.0003	0.0173
	iFLYTEK	0.0494	0.0005	0.0232
	CATL	0.0475	0.0005	0.0227
	SMIC	0.0861	0.0009	0.0306
	SUM	0.2563	0.0006	0.0236
EMH	BYD	0.0463	0.0005	0.0223
	Hikvision	0.0278	0.0003	0.0173
	iFLYTEK	0.0490	0.0005	0.0223
	CATL	0.0479	0.0005	0.0223
	SMIC	0.0825	0.0009	0.0300
	SUM	0.2535	0.0138	0.1174

The table reports the predictive accuracy of different models for individual stocks, presenting for each stock the Sum of Squared Errors (SSE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Lower values indicate stronger predictive performance.

4.3 Portfolio

We select the return data of Chinese treasury bonds (risk-free assets) from August 3, 2023, to September 30, 2024 and combine them with the returns of stocks - BYD, Hikvision, iFLYTEK, CATL, and SMIC - to construct investment portfolios. Table 8 presents two types of portfolios: one with the highest Sharpe ratio (0.038829, which

pursues high returns per unit of risk) and one with the minimum variance (0.003945, focusing on low volatility for stability). The former optimises weights to balance risk and return, while the latter focuses on reducing volatility to ensure stability. In the efficient frontier graph (Fig. 8), the horizontal axis represents volatility, and the vertical axis represents expected return. Each scatter point illustrates the risk-return relationship of a different configuration. The green star (minimum volatility) and red star (maximum Sharpe ratio) serve as typical anchor points. Investors can select suitable portfolios within the frontier curve and scatter range based on their risk preferences to achieve personalised risk and return matching.

Table 8 Results of Portfolio Decision

Portfolio Type	Stocks Name	Weights	Key Metric	Value
Highest Sharpe Ratio Portfolio	BYD	0.0382	Sharpe Ratio	0.0388
	Hikvision	0.0445		
	iFLYTEK	0.0267		
	CATL	0.5638		
	SMIC	0.3187		
	Chinese Government Bonds	0.0081		
Minimum Variance Portfolio	BYD	0.0067	Portfolio Variance	0.0039
	Hikvision	0.3247		
	iFLYTEK	0.0371		
	CATL	0.0561		
	SMIC	0.0112		
	Chinese Government Bonds	0.5642		

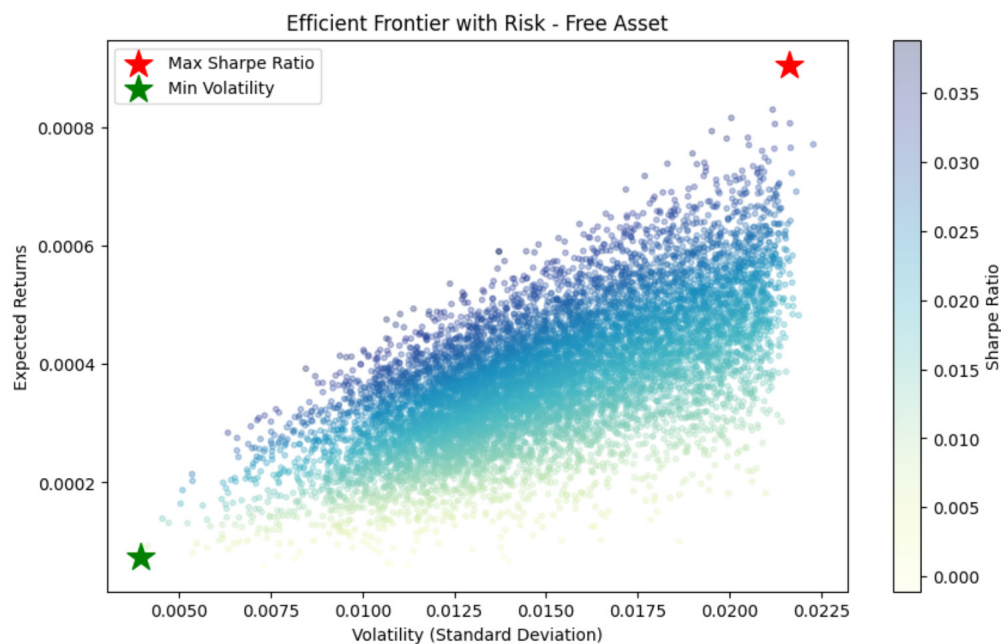


Figure 8. Efficient Frontier Graph

5. Conclusion

5.1 Research conclusions

We investigate the impact of investor sentiment extracted from Guba forum posts on the daily returns of five large-cap Chinese technology stocks from 2023 to 2024. Using pooled regressions and two-stage residual models, we find evidence suggesting that lagged sentiment variables are statistically associated with subsequent variations in re-

turns. In particular, the sentiment extracted from negative posts appears to exert a relatively more potent effect than that from positive posts, consistent with the asymmetric sentiment effect documented in the behavioural finance literature.

Forecasting exercises indicate that models incorporating sentiment proxies tend to yield lower prediction errors compared to a baseline Efficient Market Hypothesis (EMH) framework, as measured by SSE, MSE, and RSE. While these results do not necessarily imply market inef-

iciency, they suggest that sentiment-related information may contain some predictive content for short-term return movements, at least within our sample.

In a portfolio application, we construct a long-short strategy utilising sentiment signals in conjunction with Chinese government bonds. The resulting portfolio achieves a relatively higher Sharpe ratio and lower realised volatility compared to several alternative allocations considered in this study. However, it should be emphasised that this exercise is illustrative rather than conclusive, and the performance may be sensitive to transaction costs, market conditions, and model specification.

It is imperative to recognise the inherent limitations of this study. The empirical analysis is based on a relatively short time horizon and a small cross-section of technology stocks, which may limit the external validity of the findings. Moreover, the sentiment index, while constructed from real-time textual data, remains a proxy measure and may not fully capture the complexity of investor beliefs or expectations. Future research could benefit from expanding the sample size, incorporating additional control variables such as turnover and volatility, and employing more robust identification strategies such as Vector Autoregression (VAR) or structural models, to better isolate the causal effects of sentiment on returns.

5.2 Policy and Practical Recommendations

From a practical perspective, retail investors may consider treating strongly worded statements on investment forums—such as claims that “a bull market is imminent” or that “a crash is underway”—as informal contrarian indicators. Position-sizing decisions conditioned on unusually positive or negative sentiment could, in principle, limit exposure during periods of exuberance and add risk when fear appears widespread. Incorporating a publicly available sentiment gauge into one’s risk-management checklist may also reduce the likelihood of emotion-driven trades and modestly improve the timing of position changes.

We further suggest that both investors and regulators treat sentiment analytics as a supplementary tool within standard market-monitoring protocols. Given the short-lived nature of sentiment shocks and the continued dominance of market-wide return factors, regulatory initiatives should utilize sentiment data to enhance transparency and investor education, rather than targeting sentiment itself. Examples include publishing real-time “sentiment heat maps” during episodes of elevated retail leverage or abnormal social media traffic, thereby alerting participants to potential herding behaviour. Financial media platforms could also be encouraged to moderate sensational headlines and clearly distinguish between paid promotional material and independent news coverage. Collectively, these measures are expected to dampen information noise,

mitigate behavioural biases, and offer an additional layer of protection to less experienced investors.

Author Contributions

For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used: Conceptualization, Junyi He and Chenkai Zhu; methodology, Junyi He and Wangzheng Wanyan; software, Junyi He developed the two-stage residual regression, return forecasting, and descriptive-statistics modules, Wangzheng Wanyan developed the pooled-regression and portfolio-construction modules; validation, Junyi He and Wangzheng Wanyan; formal analysis, Junyi He and Chenkai Zhu; investigation, Huandong Li; resources, Huandong Li; data curation, Huandong Li; writing—original draft preparation, Chenkai Zhu; writing—review and editing, Chenkai Zhu and Junyi He; visualization, Junyi He and Wangzheng Wanyan; supervision, Chenkai Zhu; project administration, Chenkai Zhu; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Acknowledgement

Junyi He, Wangzheng Wanyan, Huandong Li, and Chenkai Zhu contributed equally to this work and should be considered co-first authors.

References

- [1] He, P., Sun, Y., Zhang, Y. and Li, T., 2021. COVID-19's impact on stock prices across different sectors—An event study based on the Chinese stock market. In *Research on pandemics* (pp. 66-80). Routledge.
- [2] Chebbi, K., Ammer, M.A. and Hameed, A., 2021. The COVID-19 pandemic and stock liquidity: Evidence from S&P 500. *The Quarterly Review of Economics and Finance*, 81, pp.134-142.
- [3] Hsu, P.H., Li, K. and Tsou, C.Y., 2023. The pollution premium. *The Journal of Finance*, 78(3), pp.1343-1392.
- [4] Pandey, D.K. and Kumari, V., 2021. Event study on the reaction of the developed and emerging stock markets to the 2019-nCoV outbreak. *International Review of Economics & Finance*, 71, pp.467-483.
- [5] Lee, W.Y., Jiang, C.X. and Indro, D.C., 2002. Stock market volatility, excess returns, and the role of investor sentiment. *Journal of banking & Finance*, 26(12), pp.2277-2299.
- [6] Baker, M. and Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4), pp.1645-1680.
- [7] Wright, C. and Swidler, S., 2023. Abnormal trading volume, news and market efficiency: Evidence from the Jamaica Stock Exchange. *Research in International Business and Finance*, 64, p.101804.
- [8] Ph, H. and Uchil, R., 2020. Influence of investor sentiment

and its antecedent on investment decision-making using partial least square technique. *Management Research Review*, 43(11), pp.1441-1459.

[9] Loewenstein, G.F., Weber, E.U., Hsee, C.K. and Welch, N., 2001. Risk as feelings. *Psychological bulletin*, 127(2), p.267.

[10] Lucey, B.M. and Dowling, M., 2005. The role of feelings in investor decision-making. *Journal of economic surveys*, 19(2), pp.211-237.

[11] See-To, E.W. and Yang, Y., 2017. Market sentiment dispersion and its effects on stock return and volatility. *Electronic Markets*, 27(3), pp.283-296.

[12] Platt, M.L. and Huettel, S.A., 2008. Risky business: the neuroeconomics of decision making under uncertainty. *Nature neuroscience*, 11(4), pp.398-403.

[13] Lam, K., Liu, T. and Wong, W.K., 2010. A pseudo-Bayesian model in financial decision making with implications to market volatility, under-and overreaction. *European Journal of Operational Research*, 203(1), pp.166-175.

[14] Baker, M. and Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4), pp.1645-1680.

[15] Jiang, F., Lee, J., Martin, X. and Zhou, G., 2019. Manager sentiment and stock returns. *Journal of Financial Economics*, 132(1), pp.126-149.

[16] Kahneman, D. and Tversky, A., 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals*

of financial decision making: Part I (pp. 99-127).

[17] Shleifer, A. and Vishny, R.W., 1997. The limits of arbitrage. *The Journal of finance*, 52(1), pp.35-55.

[18] Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3), pp.1139-1168.

[19] Zhang, K., Ren, W., Zhang, Y., & You, J. (2017). Research on stock prediction method based on information of investors' comments. *Journal of Beijing University of Posts and Telecommunications*, 42(5), 1-10. DOI: 10.16508/j.cnki.11-5866/n.2017.05.013.

[20] Silva de Carvalho, R., da Fonseca Tonin, J.M. and Raimundini Sanches, S.L., 2023. Stock Return Predictability based on Textual Sentiment Analysis: a review. *Revista de Educação e Pesquisa em Contabilidade*, 17(2).

[21] Lai, K., Chen, H., Le, G., & Dong, Y. (2014). Can Mood Predict Stock Market? *South China Journal of Psychology*, 30(007), 1770-1781. DOI: 10.3724/SP.J.1042.2014.01770.

[22] Kim, K. and Ryu, D., 2020. Predictive ability of investor sentiment for the stock market. *Romanian Journal of Economic Forecasting*, 23(4), pp.33-46.

[23] W. Zhang, X. Gong, C. Wang, and X. Ye, 'Predicting stock market volatility based on textual sentiment: A nonlinear analysis', *Journal of Forecasting*, vol. 40, no. 8, pp. 1479–1500, Dec. 2021, doi: 10.1002/for.2777.