Where Has the Foreign Signal Gone?

Read latest version here

Piyush Akimitsu*, Ulrich Hounyo**

SUNY at Albany

Abstract

Social media sentiment predicts stock losses and gains, yet for which stocks and why remains unclear. We propose that the answer lies in the match between who discusses a stock on Twitter and who holds genuine informational advantages about it. English-language Twitter discourse about U.S.-listed stocks is concentrated among U.S.-based investors, who are well-positioned to assess domestic firms through familiarity with U.S. business conditions, regulatory filings, and industry dynamics. For foreign stocks traded as American Depositary Receipts, the best-informed investors are based in those companies' home countries, holding superior knowledge of firm fundamentals and domestic market conditions, yet they contribute disproportionately less to English-language financial Twitter discussion of those specific stocks. This mismatch predicts that U.S. Twitter sentiment should be a more reliable signal for domestic stocks than for foreign ones. Using a large panel of firm-quarter observations, we confirm this: sentiment uncertainty predicts subsequent losses for domestic stocks but leaves foreign losses largely unpredicted. This predictive gap is present for the most typical stocks, widens for stocks of intermediate benchmark similarity, and collapses entirely for the least typical stocks, where domestic and foreign sentiment become equally informative about future losses and gains. Structural estimation confirms that the domestic advantage reflects genuine signal precision rather than mechanical differences in return volatility. Exploiting COVID-19 as a shock to information geography, we find that the pandemic amplified rather than equalized the domestic advantage, consistent with rising information complexity increasing the value of country-level investor knowledge. These findings link the home bias literature to the informativeness of crowd-sourced signals and suggest that the value of social media data is geographically structured.

*Email: pgade@albany.edu

**Email: khounyo@albany.edu

Keywords: Social Media Sentiment, Asset Pricing, Information Frictions, Home Bias

JEL: C22, C53, G12

1. Introduction

A growing literature examines relationships between social media sentiment and stock returns. Studies have documented correlations between Twitter activity and subsequent price movements, developed trading strategies based on sentiment signals, and applied increasingly sophisticated techniques to extract information from text. Yet a fundamental question remains underexplored: why should social media sentiment contain information about future returns, and for which stocks should this information be most relevant? Without an economic mechanism linking sentiment to returns, empirical findings risk being dismissed as data mining or spurious correlation. This paper provides such a mechanism and documents how sentiment informativeness varies systematically across domestic and foreign securities.

The economic mechanism we propose centers on the match between who discusses a stock on Twitter and who holds genuine informational advantages about it. Although Twitter is a global platform, the U.S. users accounted for approximately 20 percent of monetizable daily active users as of 2019 (Twitter, Inc., 2019). English-language financial Twitter discussion of U.S.-listed stocks is disproportionately driven by U.S.-based participants. Paydarzarnaghi and Rakowski (2023) analyze over 9.7 million stock-specific tweets and show that U.S.-demographic user groups drive equity price reactions significantly more than other groups, establishing that the price-relevant signal in English-language financial Twitter is concentrated among U.S.-based investors. These investors exhibit well-documented home bias, holding portfolios disproportionately weighted toward domestic securities (French and Poterba, 1991), which reflects informational advantages: they are familiar with U.S. business conditions, regulatory filings, and industry dynamics, and rationally deepen rather than diversify this knowledge (Van Nieuwerburgh and Veldkamp, 2009; Mondria and Wu, 2010). Under finite information-processing capacity (Sims, 2003), U.S. investors allocate relatively little attention to foreign stocks where their informational disadvantage makes processing less productive, while home-country investors of those foreign firms do the same in English-language Twitter. For foreign stocks traded as American Depositary Receipts, the best-informed investors are based in those companies' home countries. They hold superior knowledge of firm fundamentals and domestic market conditions that is not fully represented in English-language financial Twitter, which is disproportionately driven by U.S.-based investors who face a comparative disadvantage in assessing

those firms. When arbitrage is constrained (Shleifer and Vishny, 1997), this information mismatch can generate persistent differences in sentiment informativeness across security types.

This framework yields a central prediction: sentiment uncertainty should predict the components of the return distribution—downside losses and compressed upside gains—differently for domestic and foreign stocks. For domestic securities, where U.S.-based investors possess informational advantages about those firms, Twitter sentiment may capture meaningful signals about fundamentals. Sentiment volatility and dispersion among Twitter users may proxy for unresolved uncertainty that subsequently materializes in downside losses and compressed gains. For foreign securities, the price-relevant Twitter signal is generated by investors who are comparatively less informed about those firms, attenuating its predictive content. The predictive content of U.S. Twitter sentiment should therefore differ systematically across security types.

We test this prediction using a comprehensive dataset of Twitter sentiment matched to stock returns for both domestic U.S. equities and foreign stocks traded as American Depositary Receipts. Our empirical strategy distinguishes between sentiment level and sentiment uncertainty. While prior literature focuses primarily on whether bullish sentiment predicts higher returns, we examine whether disagreement and volatility in sentiment predict future risk outcomes. Sentiment volatility captures temporal instability in investor views within a quarter, while sentiment dispersion captures cross-sectional disagreement among Twitter users on any given day. Both measures may proxy for underlying uncertainty about fundamentals, but through different channels.

Our identification strategy addresses the central concern that sentiment may simply react to past returns rather than predict future returns. We employ several complementary approaches. All specifications use lagged sentiment to predict subsequent outcomes, ensuring temporal precedence. We exploit cross-sectional heterogeneity: if sentiment merely reacted to prices, its predictive power should be uniform across stocks. Our theoretical framework predicts differential effects by benchmark similarity and foreign status, and finding these patterns supports a causal interpretation. We further show that the domestic-foreign differential exhibits a systematic pattern across stocks' similarity to the global cross-sectional benchmark—widening for medium-similarity stocks before collapsing for the least similar stocks—and survives controls for macroeconomic uncertainty, consistent with an information frictions mechanism rather than spurious correlation.

The empirical findings confirm the predictions of the theoretical framework. Sentiment volatil-

ity predicts subsequent losses for high-similarity stocks: stocks whose Twitter sentiment fluctuates most within a quarter go on to suffer substantially larger losses the following quarter, and this relationship is far stronger for domestic stocks than for foreign stocks. For foreign stocks, the relationship is attenuated or reversed, consistent with U.S. Twitter sentiment capturing noise relative to the superior information held by home-country investors of those foreign firms. Sentiment dispersion exhibits a different pattern: it predicts compressed future gains for low-similarity stocks, with high-dispersion stocks experiencing approximately 15 to 18 percentage points less in next-quarter total gains than low-dispersion stocks. These heterogeneous effects across security types and sentiment measures provide support for the information frictions mechanism.

Our analysis makes several contributions to the literature. We move beyond documenting sentiment-return correlations to identifying the economic mechanism that generates these correlations. The mechanism we propose, grounded in the literatures on incomplete information (Merton, 1987), attention-driven trading (Barber and Odean, 2008), information immobility (Van Nieuwerburgh and Veldkamp, 2009), and rational inattention (Sims, 2003), generates testable predictions that our empirical analysis confirms.[1] We distinguish between sentiment level and sentiment uncertainty, showing that the informational content of social media lies not only in its direction but also in its stability and consensus. We document systematic heterogeneity in predictive relationships across domestic and foreign securities, providing evidence that sentiment effects reflect information frictions rather than behavioral biases that should affect all stocks uniformly. The remainder of the paper proceeds as follows. Section 2 briefly describes the data and key variables. Section 3 presents reduced-form evidence from quintile portfolio analysis, documenting that sentiment uncertainty predicts future losses and gain outcomes differentially across domestic and foreign stocks. Section 4 reviews the related literature and develops the theoretical framework explaining these patterns. Section 5 presents structural estimation of sentiment precision, directly testing the model's central prediction. Section 6 reports robustness checks, including a COVID-19 identification analysis that exploits the pandemic as a shock to information geography. Section 7 concludes. Appendix A

[1] Christopher A. Sims (1942–2026), who pioneered the rational inattention framework, passed away on March 14, 2026. His insight that agents with finite information-processing capacity optimally allocate attention to where informational returns are highest provides the microfoundation for the geographic structure of social-media informativeness documented here.

provides detailed data construction and methodology.

2. Data

This section briefly describes the data sources and key variables. Detailed data construction methodology appears in Appendix A.

2.1. Sample and Sources

Our analysis combines three data sources, namely stock returns from CRSP-Compustat, Twitter sentiment from Context Analytics, and macroeconomic controls from standard sources. The final sample contains 186,276 firm-quarter observations covering 7,379 stocks from 2012Q1 through 2024Q4, comprising 153,328 domestic observations (82.3%) and 32,948 foreign observations (17.7%).

Stock Returns. We construct quarterly excess returns and decompose them into gains and losses. Total losses (`total_loss_excess_ret`) capture cumulative negative returns within each quarter, while total gains (`total_gain_excess_ret`) capture cumulative positive returns. All outcome variables are winsorized at the 1st and 99th percentiles.[2]

Twitter Sentiment. Context Analytics provides daily Raw-S sentiment scores based on tweets mentioning individual stocks, where Raw-S is a standardized measure of the net sentiment expressed in each tweet. Although Twitter's global user base is predominantly international (Twitter, Inc., 2019), English-language financial Twitter discussion of U.S.-listed stocks is disproportionately driven by U.S.-based investors (Paydarzarnaghi and Rakowski, 2023), which motivates our focus on the domestic-foreign differential in sentiment informativeness. We aggregate daily data to quarterly frequency to construct two measures of sentiment uncertainty. Sentiment volatility (`s_sd`) is the within-quarter standard deviation of daily sentiment scores, capturing temporal instability. Sentiment dispersion (`s_dispersion_mean`) is the quarterly mean of Context Analytics' daily source diversity metric, capturing cross-sectional disagreement among Twitter users. Context Analytics implemented a methodology change in 2017, transitioning from cashtag-only tweet identification (tweets referencing a stock's ticker symbol preceded by $, e.g., $AAPL) to a topic model approach,

[2]Formally, let $r^e_{id} = r_{id} - r^f_d$ denote the daily excess return for stock i on day d, where r_{id} is the CRSP daily return (`DlyRet`) and r^f_d is the daily risk-free rate from the Fama-French data library. Then $\texttt{total_loss_excess_ret}_{it} = \sum_{d \in q_t} \min(r^e_{id}, 0)$ and $\texttt{total_gain_excess_ret}_{it} = \sum_{d \in q_t} \max(r^e_{id}, 0)$, where the sum is over all trading days d in quarter t. By construction, `total_loss_excess_ret` ≤ 0 and `total_gain_excess_ret` ≥ 0.

which reduces sentiment levels but should not affect cross-sectional comparisons within periods. We use post-2017 data as the primary sample and report pre-2017 robustness checks.

Similarity Index. We construct a measure of how closely each stock's characteristics resemble the typical stock in our sample using Random Forest feature importance weights. First, we estimate a Random Forest model predicting quarterly excess returns using 72 predetermined firm characteristics (size, Fama-French factor betas, lagged volatility, liquidity, and valuation ratios). The out-of-bag R^2 is 4.28%. The weights w_k are the normalized Gini-based feature importances from the Random Forest, which sum to 1 by construction in scikit-learn ($\sum_{k=1}^{K} w_k = 1$). Second, for each stock i in quarter t, we compute the weighted cosine similarity between its standardized feature vector $\mathbf{x}_{i,t}$ and a global benchmark vector $\mathbf{b}_t$, defined as the median of each feature across *all stocks*, both domestic and foreign, in that quarter:

$$\text{similarity}_{i,t} = \frac{\sum_{k=1}^{K} w_k \, x_{i,t,k} \, b_{t,k}}{\sqrt{\sum_{k=1}^{K} w_k \, x_{i,t,k}^2} \sqrt{\sum_{k=1}^{K} w_k \, b_{t,k}^2}} \tag{1}$$

where $w_k \geq 0$ are the normalized Random Forest feature importance weights satisfying $\sum_{k=1}^{K} w_k = 1$, and $K = 72$ is the number of features. The global benchmark captures the "typical" stock in our full sample each quarter. The resulting distribution is heavily left-skewed (mean 0.90, median 0.97), as shown in Figure I. We partition stocks into three strata based on this distribution: *high-similarity* stocks ($\phi \geq 0.98$; 38.5% of observations), which closely match the global cross-sectional benchmark in firm characteristics and represent the most typical stocks in the full sample; *medium-similarity* stocks ($0.90 \leq \phi < 0.98$; 42.1%), which share many characteristics with the benchmark but exhibit some idiosyncratic features; and *low-similarity* stocks ($\phi < 0.90$; 19.5%), which have distinctly atypical characteristics. As we document in Section 3.8, the predictive content of sentiment for domestic stocks strengthens systematically as similarity decreases, with the domestic-foreign differential exhibiting a non-monotonic pattern.

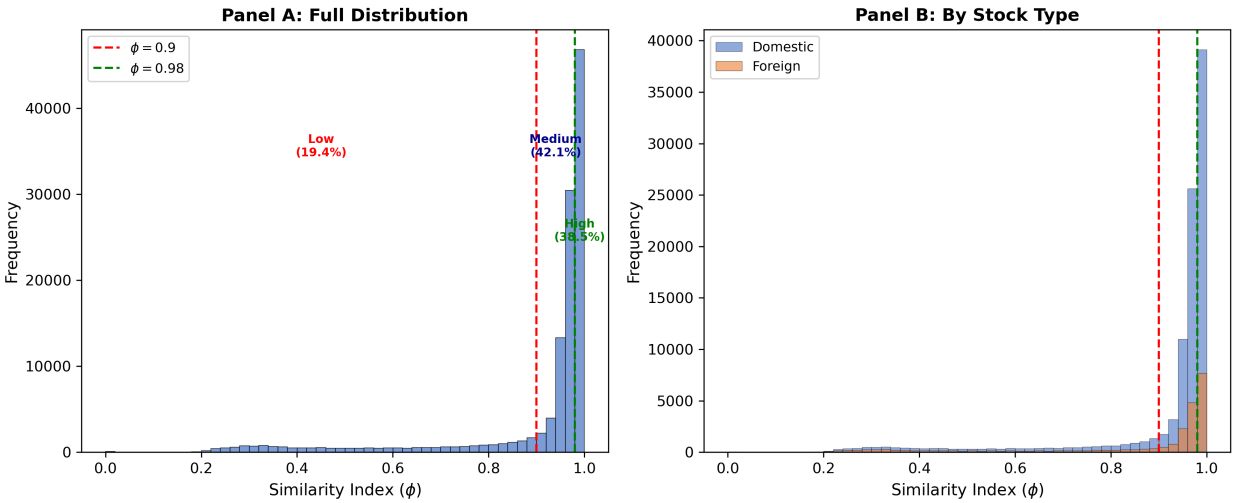


FIGURE I
Distribution of Similarity Index

Notes: Panel A shows the distribution of the similarity index ϕ across all firm-quarters in the post-2017 sample ($N = 126{,}095$). Dashed lines indicate the cutoffs used to define the three similarity strata: $\phi = 0.90$ (red) and $\phi = 0.98$ (green). The distribution is heavily left-skewed with the majority of observations concentrated near 1.0. Panel B shows the distribution separately for domestic and foreign stocks.

2.2. Summary Statistics

Table I presents summary statistics for key variables by stock type and similarity stratum.

Table I: Summary Statistics

	High ($\phi \geq 0.98$)		Medium ($0.90 \leq \phi < 0.98$)		Low ($\phi < 0.90$)	
Variable	Domestic	Foreign	Domestic	Foreign	Domestic	Foreign
Observations	39,859	8,653	43,419	9,631	17,109	7,424
Mean ϕ	0.993	0.991	0.961	0.957	0.584	0.636
Sentiment volatility (s_{sd})	0.121	0.118	0.087	0.088	0.107	0.105
Sentiment dispersion	0.927	0.926	0.928	0.924	0.919	0.912
Outcome variables (winsorized at 1%/99%)						
Total losses: mean	−0.8235	−0.8856	−1.0029	−1.0170	−1.4907	−1.4986
SD	(0.4400)	(0.4995)	(0.4605)	(0.5119)	(0.5880)	(0.6454)
Total gains: mean	0.4173	0.4211	0.4385	0.4540	0.9696	0.9049
SD	(0.3368)	(0.3729)	(0.3838)	(0.4050)	(0.5308)	(0.5777)
Deviation: mean	−0.4331	−0.4922	−0.6040	−0.5987	−0.5481	−0.6201
SD	(0.4419)	(0.4944)	(0.5860)	(0.5844)	(0.6612)	(0.6610)
Return volatility: mean	0.0256	0.0269	0.0272	0.0289	0.0544	0.0556
SD	(0.0158)	(0.0189)	(0.0158)	(0.0190)	(0.0256)	(0.0311)

Notes: Sample: post-2017 period (2017Q1–2024Q4) with non-missing similarity index ($N = 126{,}095$). High similarity: $\phi \geq 0.98$. Medium similarity: $0.90 \leq \phi < 0.98$. Low similarity: $\phi < 0.90$. Sentiment volatility and dispersion are means across the raw sample. Outcome variable statistics (means and standard deviations in parentheses) use the same raw sample; each variable is computed on its own non-missing observations. Estimation samples used in the regression analyses are smaller due to analysis-specific requirements for non-missing next-quarter outcomes; see Section 3.2.

Detailed data construction methodology, including Random Forest feature importance estimation and sentiment aggregation procedures, appears in Appendix A.

3. Reduced-Form Evidence

This section presents the main empirical findings. We document two central facts. First, sentiment volatility predicts subsequent losses for high-similarity domestic stocks. Second, sentiment dispersion predicts compressed gains for low-similarity stocks. These patterns motivate the theoretical framework in Section 4 and the structural estimation in Section 5.

3.1. Overview

A key contribution of our analysis is distinguishing between sentiment *level* (bullish versus bearish) and sentiment *uncertainty* (volatility and dispersion). While prior literature focuses on whether stocks with bullish sentiment outperform those with bearish sentiment, we examine whether disagreement and volatility in sentiment predict future risk outcomes.

We construct two distinct measures of sentiment uncertainty from the Context Analytics Raw-S data. Sentiment volatility (`s_sd`) is the standard deviation of daily Raw-S scores within each stock-quarter:

$$\texttt{s_sd}_{i,t} = \sqrt{\frac{1}{D_{i,t}-1}\sum_{d=1}^{D_{i,t}}\left(\text{Raw-S}_{i,d} - \overline{\text{Raw-S}}_{i,t}\right)^2} \tag{2}$$

where $D_{i,t}$ is the number of trading days with sentiment data for stock i in quarter t, $\text{Raw-S}_{i,d}$ is the daily sentiment score, and $\overline{\text{Raw-S}}_{i,t}$ is the quarterly mean. This measure captures temporal instability in investor views: high values indicate sentiment that fluctuates dramatically across days within the quarter.

Sentiment dispersion (`s_dispersion_mean`) is the quarterly average of Context Analytics' daily S_DISPERSION variable, which measures diversity of sentiment sources on each trading day:

$$\texttt{s_dispersion_mean}_{i,t} = \frac{1}{D_{i,t}}\sum_{d=1}^{D_{i,t}}\text{S_DISPERSION}_{i,d} \tag{3}$$

where S_DISPERSION is a Context Analytics metric bounded between 0 and 1 that captures how dispersed sentiment sources are on a given day. Higher values indicate that sentiment about a

stock comes from a more diverse set of Twitter users rather than being concentrated among a few sources. This measure proxies for the breadth of investor attention and disagreement, with high dispersion indicating that many different users are expressing views about the stock, potentially with divergent opinions.

Both measures proxy for underlying uncertainty about fundamentals, but through different channels. Sentiment volatility reflects changing consensus over time, while sentiment dispersion reflects contemporaneous disagreement. To reduce the influence of extreme observations on our estimates, all outcome variables are winsorized at the 1st and 99th percentiles prior to analysis.

3.2. Portfolio Construction and Sample Stratification

For each quarter t, we sort stocks into quintiles based on lagged sentiment characteristics measured in quarter $t-1$ and examine outcomes in quarter t. This timing ensures our analysis captures predictive relationships rather than contemporaneous correlations. Quintile 1 (Q1) contains stocks with the lowest sentiment volatility or dispersion, while Quintile 5 (Q5) contains stocks with the highest sentiment volatility or dispersion. Throughout the analysis, the Q5–Q1 spread refers to the difference in mean outcomes between the highest and lowest quintile portfolios; a negative spread indicates that high-sentiment stocks perform worse than low-sentiment stocks in the following quarter.

We stratify the analysis by stock similarity to the global cross-sectional benchmark (Figure I). We begin by presenting detailed results for the two extreme strata, where the theoretical predictions are sharpest: high-similarity stocks ($\phi \geq 0.98$; 38.5% of the post-2017 sample) closely match the global median across firm characteristics and typically comprise large-cap, liquid securities; low-similarity stocks ($\phi < 0.90$; 19.5%) exhibit idiosyncratic return patterns and may have greater information asymmetry. In Section 3.8, we add the medium-similarity stratum ($0.90 \leq \phi < 0.98$; 42.1%) and show that the pattern across strata is systematic—with the domestic loss-prediction gradient strengthening as similarity decreases and the domestic-foreign differential widening before collapsing—confirming a smooth gradient rather than an artifact of extreme-group comparisons.

We examine three primary outcome variables: deviation from market (`deviation_quarterly`),[3]

[3]Formally, $\texttt{deviation_quarterly}_{it} = \sum_{d \in q_t} r^e_{id} - \sum_{d \in q_t} r^{\text{SP500}}_d$, where r^{SP500}_d is the daily S&P 500 return (`sprtrn`

measuring idiosyncratic performance; total losses (`total_loss_excess_ret`), capturing downside risk exposure; and total gains (`total_gain_excess_ret`), capturing upside potential. We also examine return volatility (`sd_excess_ret`),[4] the within-quarter standard deviation of daily excess returns.

3.3. Predictive Regressions

Table II reports R^2 values from univariate regressions of next-quarter outcomes on lagged sentiment characteristics, stratified by benchmark similarity and domestic versus foreign status. These regressions establish the baseline predictive power of each sentiment measure before turning to the quintile analysis.

Table II: Predictive Power of Lagged Sentiment Uncertainty for Next-Quarter Stock Outcomes

		High Similarity ($\phi \geq 0.98$)		Low Similarity ($\phi < 0.90$)	
Predictor	Outcome	Domestic	Foreign	Domestic	Foreign
Panel A: Sentiment Volatility (s_sd)					
	Deviation from Market	4.86%	**7.69%**	4.28%	3.28%
	Total Losses	5.43%	**8.52%**	6.52%	2.68%
	Return Volatility	1.55%	1.85%	0.54%	0.06%
Panel B: Sentiment Dispersion (s_dispersion_mean)					
	Deviation from Market	0.81%	0.53%	**1.46%**	0.68%
	Total Gains	0.01%	0.12%	0.74%	**1.07%**
	Return Volatility	0.31%	0.02%	0.07%	0.26%

Notes: This table reports R^2 values from univariate regressions of next-quarter outcomes on lagged sentiment measures. Bold values indicate the strongest effects within each panel. High Similarity: $\phi \geq 0.98$; Low Similarity: $\phi < 0.90$. Medium-similarity results ($0.90 \leq \phi < 0.98$) are reported in Table IV. All specifications use lagged sentiment (quarter $t-1$) to predict outcomes (quarter t). Outcome variables are winsorized at 1%/99%.

Panel A of Table II reveals that sentiment volatility explains 5–8% of variation in next-quarter losses and market deviation for high-similarity stocks. These R^2 values are substantial relative to typical cross-sectional return predictability in the literature. The effect is strongest for foreign stocks in the high-similarity sample, where sentiment volatility explains 7.69% of deviation and 8.52% of total losses. This apparent tension with the quintile evidence reflects the difference in estimands: univariate R^2 captures linear fit across the full distribution of outcomes including

from CRSP). This is the cumulative within-quarter excess return of the stock minus the cumulative S&P 500 return, measuring idiosyncratic quarterly performance net of broad market movements.

[4] Formally, $\texttt{sd_excess_ret}_{it} = \left[\frac{1}{D_{it}-1}\sum_{d=1}^{D_{it}}(r^e_{id} - \bar{r}^e_{it})^2\right]^{1/2}$, where D_{it} is the number of trading days with non-missing return data in quarter t and $\bar{r}^e_{it}$ is the within-quarter mean of daily excess returns. This is the realized intra-quarter volatility of the stock's daily excess returns.

outliers, while quintile spreads isolate tail behavior; the two are not contradictory, and the main results throughout use quintile spreads as the primary evidence.

Panel B demonstrates that sentiment dispersion has minimal predictive power for high-similarity stocks but predicts gains for low-similarity stocks. For high-similarity stocks, disagreement among Twitter users adds little incremental information. For low-similarity stocks with greater information asymmetry, disagreement captures meaningful uncertainty about fundamentals.

Given the 2017 methodology change documented in Section 2, we conduct the primary quintile analysis using the post-2017 sample (2017Q1–2024Q4), which employs consistent topic model methodology throughout. The raw post-2017 sample comprises 126,095 firm-quarter observations with non-missing similarity index across the three strata: 48,512 high-similarity (82.2% domestic), 53,050 medium-similarity (81.8% domestic), and 24,533 low-similarity (69.7% domestic). Estimation samples are smaller because each analysis requires non-missing values only for the outcome variable it uses: the sentiment volatility analysis uses 42,950 high-similarity, 47,552 medium-similarity, and 20,973 low-similarity observations, while the sentiment dispersion analysis uses 40,928, 45,228, and 21,052 observations, respectively. Sample sizes differ within each stratum across the two analyses because observations missing next-quarter losses but not next-quarter gains are included in the dispersion analysis but not the volatility analysis.

Quintile assignments are balanced, with each quintile containing approximately 20% of observations within each quarter. We begin with the extreme strata below and extend to the full distribution in Section 3.8. We present pre-2017 results as robustness checks in Section 6.

3.4. Sentiment Volatility and Downside Risk

Figure II presents the quintile analysis for sentiment volatility using the high-similarity sample. Stocks are sorted into quintiles based on lagged `s_sd`, with Q1 containing stocks with the most stable sentiment and Q5 containing stocks with the most volatile sentiment.

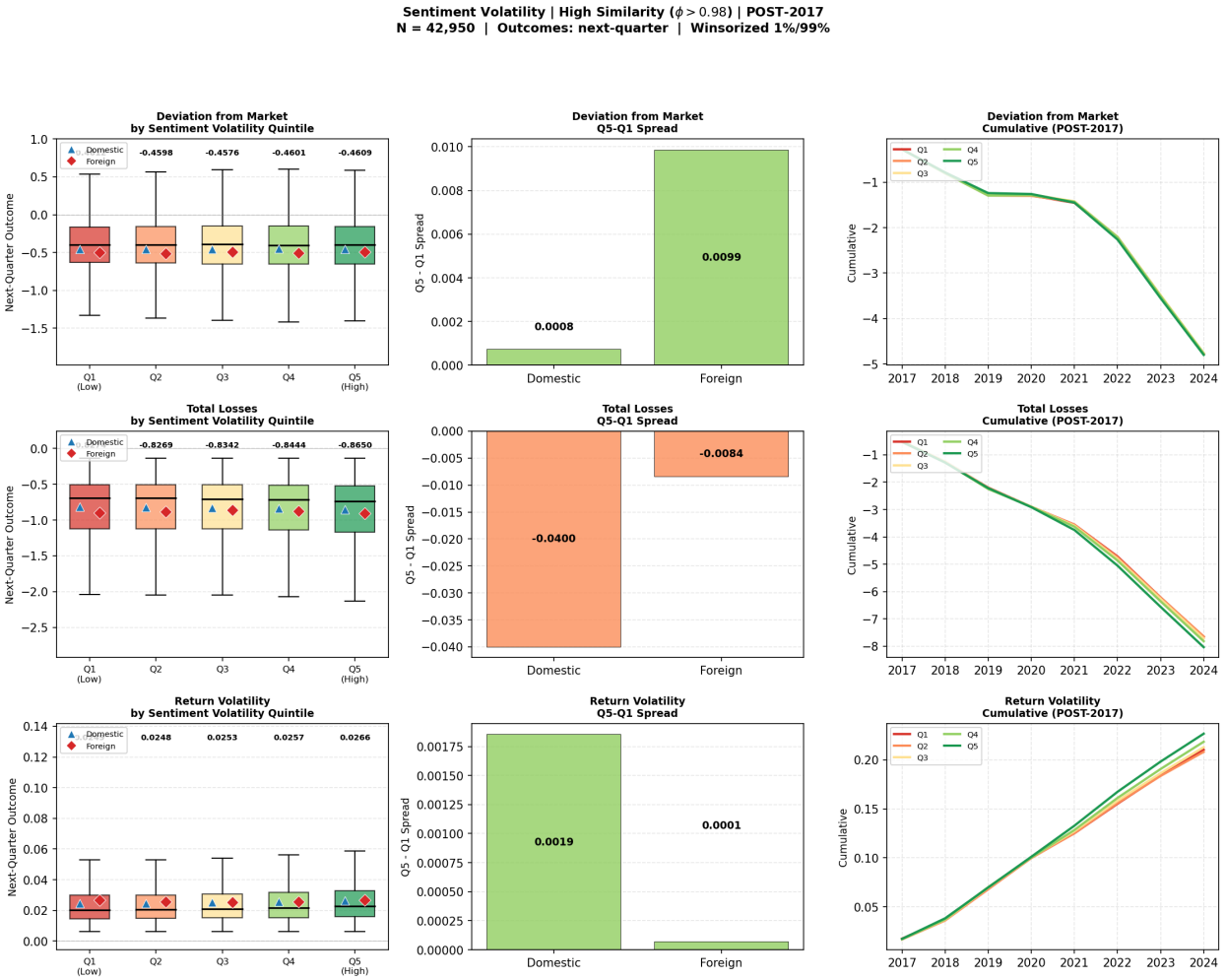


FIGURE II
Sentiment Volatility and Next-Quarter Outcomes: High Similarity Sample (Post-2017)

Notes: Stocks sorted into quintiles based on lagged sentiment volatility (`s_sd`). Sample restricted to similarity index ≥ 0.98 and post-2017 period ($N = 42,950$). Q1 = lowest sentiment volatility; Q5 = highest sentiment volatility. Left column: box plots of next-quarter outcomes by quintile with domestic means (blue triangles) and foreign means (red diamonds). Middle column: Q5–Q1 spread for domestic and foreign stocks separately. Right column: cumulative value by quintile, 2017–2024. Outcome variables winsorized at 1%/99%.

The top row of Figure II presents results for deviation from market. Quintile means range from −0.458 to −0.461, indicating relatively flat average deviations across sentiment volatility quintiles. The Q5–Q1 spreads are small and positive for both domestic (+0.0008) and foreign stocks (+0.0099), suggesting that in the post-2017 period, sentiment volatility has limited predictive power for market deviations in the high-similarity sample.

The middle row presents results for total losses. Both domestic and foreign stocks with high sentiment volatility experience larger subsequent losses, with Q5–Q1 spreads of −0.0400 for domes-

tic stocks and −0.0084 for foreign stocks. The domestic spread is economically meaningful, with stocks in the highest sentiment volatility quintile experiencing 4 percentage points more in total losses than those in the lowest quintile. The right panel shows cumulative losses over the 2017–2024 period, with Q5 stocks consistently accumulating the largest losses.

The bottom row examines return volatility. Higher sentiment volatility predicts modestly higher return volatility for domestic stocks (spread of +0.0019), while the effect for foreign stocks is negligible (+0.0001). These effects are small compared to the loss predictions.

The central finding is that for high-similarity stocks, sentiment volatility predicts future losses. When Twitter sentiment about a stock fluctuates dramatically within a quarter, this instability forecasts larger realized losses in the subsequent quarter, with domestic stocks showing a stronger effect. This finding is consistent with sentiment volatility proxying for unresolved information uncertainty that subsequently materializes as downside losses.

We next examine whether sentiment dispersion—a cross-sectional measure of disagreement among Twitter users—carries similar predictive content, and whether that content is concentrated among the same stock types.

3.5. Sentiment Dispersion and Low-Similarity Stocks

Figure III presents the quintile analysis for sentiment dispersion using the low-similarity sample. Stocks are sorted based on lagged `s_dispersion_mean`, with Q1 containing stocks with the most agreement among Twitter users and Q5 containing stocks with the most disagreement.

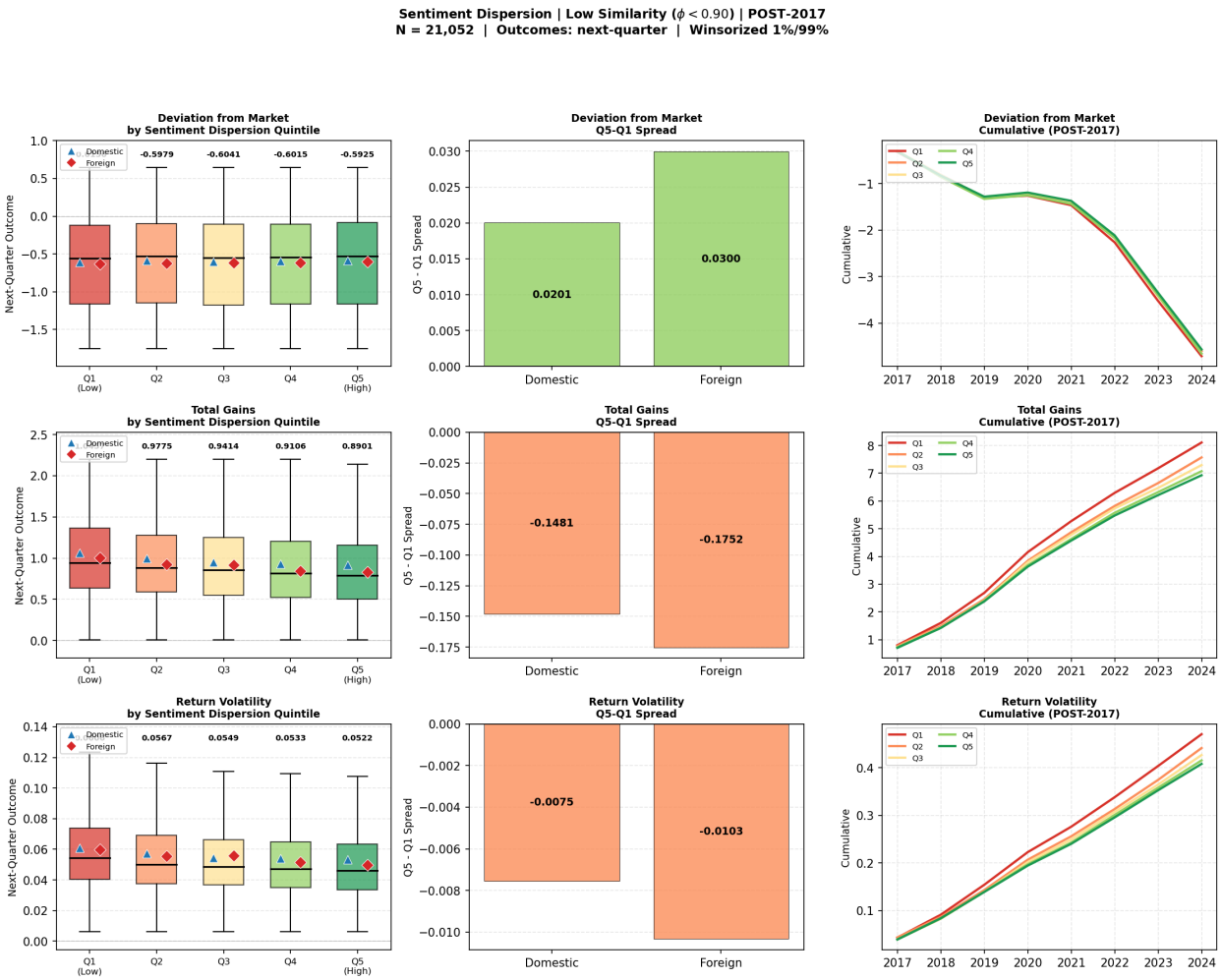


FIGURE III
Sentiment Dispersion and Next-Quarter Outcomes: Low Similarity Sample (Post-2017)

Notes: Stocks sorted into quintiles based on lagged sentiment dispersion (`s_dispersion_mean`). Sample restricted to similarity index < 0.90 and post-2017 period ($N = 21,052$). Q1 = lowest dispersion (high agreement); Q5 = highest dispersion (high disagreement). Left column: box plots with domestic means (blue triangles) and foreign means (red diamonds). Middle column: Q5–Q1 spread by stock type. Right column: cumulative value by quintile, 2017–2024. Outcome variables winsorized at 1%/99%.

The top row of Figure III shows results for deviation from market. Both domestic and foreign stocks exhibit positive Q5–Q1 spreads (domestic: +0.0201; foreign: +0.0300), indicating that stocks with high sentiment dispersion subsequently deviate more positively from the market. Notably, the effect is stronger for foreign stocks, suggesting that disagreement captures more unresolved information for low-similarity foreign stocks.

The middle row presents the most striking finding. Both domestic and foreign stocks show large negative Q5–Q1 spreads for total gains (domestic: −0.1481; foreign: −0.1752). High sentiment dispersion predicts substantially lower future gains. When Twitter users disagree about a stock,

subsequent positive returns are compressed. The right panel illustrates this over the 2017–2024 period, with Q1 stocks (low dispersion, high agreement) consistently accumulating substantially greater gains than Q5 stocks (high dispersion, high disagreement). The economic magnitude is considerable: stocks with the highest disagreement experience approximately 15–18 percentage points less in total gains than stocks with the most agreement.

The bottom row examines return volatility. Higher sentiment dispersion predicts lower return volatility for both stock types (domestic: -0.0075; foreign: -0.0103). This counterintuitive finding suggests that disagreement in sentiment may precede periods of price stability rather than volatility, perhaps because disagreement prevents the coordination necessary for large directional moves.

The central finding is that for low-similarity stocks, sentiment dispersion strongly predicts lower future gains. When Twitter users disagree about a stock's prospects, this disagreement forecasts compressed upside returns in the subsequent quarter. The consistency of this pattern across both domestic and foreign stocks, and, as we show in Section 6, across both methodology periods, strengthens confidence that the effect reflects genuine information content rather than measurement artifacts.

Section 3.8 extends both analyses to the medium-similarity stratum, revealing a non-monotonic gradient in the domestic-foreign differential across the full similarity distribution.

3.6. Summary of Reduced-Form Findings

The reduced-form evidence establishes two key facts that motivate our theoretical framework:

1. **Fact 1: Sentiment volatility predicts losses for domestic stocks.** In the high-similarity sample, stocks with volatile Twitter sentiment experience substantially larger losses in the following quarter, with domestic stocks showing a Q5–Q1 spread of -0.0400 compared to -0.0084 for foreign stocks. This asymmetry suggests that U.S. Twitter sentiment is more informative for domestic securities.
2. **Fact 2: Sentiment dispersion predicts compressed gains.** In the low-similarity sample, stocks with high disagreement among Twitter users experience approximately 15–18 percentage points less in subsequent gains. This effect is strong for both domestic and foreign stocks.

These patterns require explanation. Why does sentiment uncertainty predict losses and compressed gains differently across stock types? Why is the domestic-foreign differential stronger for

volatility than dispersion? The critical question is whether the differential patterns reflect genuine differences in sentiment *precision* (information content) or merely differences in return volatility. The theoretical framework in Section 4 addresses the first two questions; the structural estimation in Section 5 addresses the third.

3.7. Factor-Adjusted Portfolio Returns

A concern with the quintile analysis is that the Q5–Q1 spreads may reflect compensation for exposure to known risk factors rather than sentiment informativeness. To address this, we regress equal-weighted quarterly portfolio returns for each quintile on the Fama-French five factors (Mkt−RF, SMB, HML, RMW, CMA). The intercept α captures abnormal returns unexplained by factor exposures. This adjustment is particularly relevant for the domestic-foreign comparison: if domestic and foreign stocks systematically differ in their loadings on Mkt−RF, SMB, or HML—which they do, given that ADRs tend to be larger and more liquid than the average domestic stock—raw quintile spreads could reflect differential factor exposure rather than differential sentiment informativeness. The factor-adjusted α isolates the sentiment signal from these mechanical differences. Table III reports the results.

Table III: Factor-Adjusted Quintile Portfolio Returns

	Q1	Q2	Q3	Q4	Q5	Q5−Q1
High-Sim: s_sd → Losses						
Domestic α	+0.0039	−0.0001	−0.0027	−0.0073*	−0.0071	−0.0110
	(0.77)	(-0.01)	(-0.56)	(-1.83)	(-1.18)	(-1.55)
Foreign α	−0.0183	−0.0191**	−0.0218***	−0.0054	−0.0090	+0.0092
	(-1.60)	(-2.35)	(-2.58)	(-0.62)	(-0.96)	(0.66)
Medium-Sim: s_sd → Losses						
Domestic α	−0.0038	−0.0040	−0.0080**	−0.0122**	−0.0066	−0.0028
	(-0.50)	(-0.70)	(-2.30)	(-2.03)	(-1.19)	(-0.32)
Foreign α	−0.0077	−0.0159	−0.0313***	−0.0237**	−0.0063	+0.0014
	(-0.74)	(-1.56)	(-3.07)	(-2.31)	(-0.60)	(0.08)
Low-Sim: s_sd → Losses						
Domestic α	−0.0183	−0.0562***	−0.0315**	−0.0516**	−0.0507***	−0.0324**
	(-1.29)	(-4.26)	(-1.97)	(-2.57)	(-2.59)	(-2.10)
Foreign α	−0.0577***	−0.0594***	−0.0750***	−0.0551***	−0.1014***	−0.0438**
	(-4.30)	(-4.58)	(-4.47)	(-3.72)	(-4.46)	(-2.38)
Low-Sim: Dispersion → Gains						
Domestic α	−0.0589***	−0.0490***	−0.0400***	−0.0381**	−0.0208	+0.0381**
	(-2.67)	(-2.97)	(-2.72)	(-2.56)	(-1.52)	(2.07)
Foreign α	−0.1016***	−0.0536***	−0.0727***	−0.0566***	−0.0623***	+0.0394
	(-4.37)	(-3.11)	(-4.62)	(-3.35)	(-4.71)	(1.64)

Notes: Each quarter, stocks are sorted into quintiles based on lagged sentiment measures. Equal-weighted quarterly portfolio returns are regressed on Fama-French five factors (Mkt−RF, SMB, HML, RMW, CMA). α is the intercept (quarterly). Q5−Q1 is the long-short spread portfolio α. HC1 robust t-statistics in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. Post-2017 sample (2017Q1–2024Q4).

The factor-adjusted results are consistent with the information frictions interpretation, though the evidence is strongest for the low-similarity stratum. For the high-similarity stratum (sentiment volatility), the domestic Q5–Q1 spread alpha is −1.10 percentage points per quarter ($t = -1.55$), and the foreign spread alpha is +0.92 pp ($t = 0.66$); neither achieves conventional significance. The absence of a significant domestic alpha in this stratum after factor adjustment indicates that the raw quintile spread for high-similarity stocks largely reflects factor exposures rather than pure abnormal returns.

The gradient across strata clarifies this pattern. For medium-similarity stocks, both the domestic spread alpha (−0.28 pp, $t = -0.32$) and the foreign spread alpha (+0.14 pp, $t = 0.08$) are statistically indistinguishable from zero after factor adjustment. For low-similarity stocks, by contrast, both domestic (−3.24 pp, $t = -2.10$, $p < 0.05$) and foreign (−4.38 pp, $t = -2.38$, $p < 0.05$) spread alphas are significant, with the foreign alpha larger in absolute value. For low-similarity stocks, loss-sorted portfolio returns survive factor adjustment, confirming that the loss prediction

documented in Section 3.4 reflects genuine abnormal returns. The dispersion-sorted gains panel (Section 3.5) yields a Q5-Q1 factor-adjusted spread of +3.81 pp (domestic, $t = 2.07$) and +3.94 pp (foreign, $t = 1.64$), driven by Q1 (low-dispersion) stocks generating more negative factor-adjusted alphas than Q5 stocks. For atypical low-similarity stocks that are poorly described by standard factor models, we treat the raw dispersion-gains result as the more economically meaningful characterization.

To contextualize the loss-prediction magnitudes: for low-similarity stocks, the domestic loss-sorted spread alpha of −3.24 pp per quarter corresponds to approximately −13% annualized, while the foreign spread of −4.38 pp implies approximately −18% annualized. These are economically substantial, comparable to well-documented anomalies such as short-term reversal (∼5–8% annualized; Fama and French 2012) and post-earnings announcement drift. For high- and medium-similarity stocks, the loss-sorted factor-adjusted alphas are insignificant, suggesting that the sentiment volatility signal in those strata is captured by known risk premia rather than representing a distinct source of return predictability.

3.8. Similarity Gradient: The Medium-Similarity Stratum

The preceding analysis focuses on the two extreme strata of the similarity distribution. A natural concern is whether the documented patterns reflect cherry-picked subsamples. To address this, we extend the quintile analysis to include the medium-similarity stratum ($0.90 \leq \phi < 0.98$), which contains 42.1% of the raw post-2017 sample (53,050 observations; 47,552 in the sentiment volatility estimation sample). Table IV and Figure IV report Q5–Q1 spreads across all three strata.

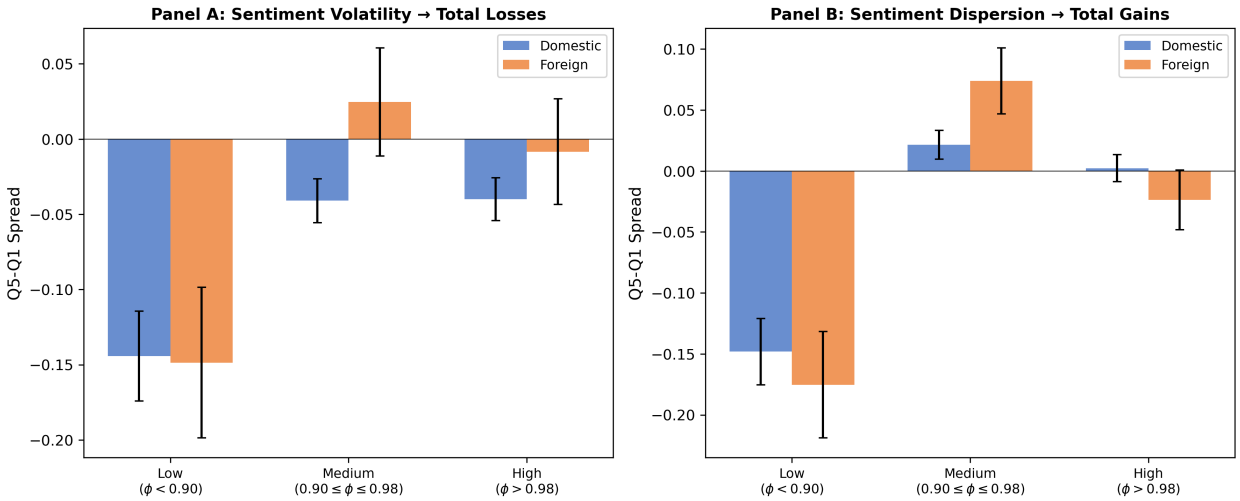


FIGURE IV
Q5–Q1 Quintile Spreads Across Similarity Strata

Notes: Bars show Q5–Q1 quintile spreads for domestic (blue) and foreign (orange) stocks across three similarity strata. Error bars represent 95% confidence intervals from Welch t-tests. Panel A: sentiment volatility (s_{sd}) predicting next-quarter total losses. Panel B: sentiment dispersion predicting next-quarter total gains. Post-2017 sample. Sample sizes across all three strata: $N = 111{,}475$ for the sentiment volatility (Panel A) analysis and $N = 107{,}208$ for the sentiment dispersion (Panel B) analysis; see Section 3.2 for stratum-specific counts.

Table IV: Quintile Spreads Across Similarity Strata

	High ($\phi > 0.98$)		Medium ($0.90 \leq \phi \leq 0.98$)		Low ($\phi < 0.90$)	
	Domestic	Foreign	Domestic	Foreign	Domestic	Foreign
Panel A: Sentiment Volatility (s_{sd}) → Next-Quarter Total Losses						
Q5–Q1 Spread	-0.0400***	-0.0084	-0.0410***	0.0247	-0.1443***	-0.1487***
	(0.0073)	(0.0179)	(0.0074)	(0.0184)	(0.0153)	(0.0255)
Panel B: Sentiment Dispersion → Next-Quarter Total Gains						
Q5–Q1 Spread	0.0023	-0.0238*	0.0215***	0.0739***	-0.1481***	-0.1752***
	(0.0057)	(0.0125)	(0.0060)	(0.0138)	(0.0139)	(0.0223)
Panel C: Continuous Interaction ($s_{sd} \times \phi$) → Next-Quarter Total Losses, pooled across all strata						
	Domestic (all strata)			*Foreign (all strata)*		
$\hat{\beta}(s_{sd} \times \phi)$	+0.0053***			+0.0092**		
	(0.0019); $N = 97{,}173$			(0.0039); $N = 22{,}175$		

Panels A and B report Q5–Q1 quintile spreads from portfolio sorts within each quarter. Standard errors (in parentheses) are Welch t-test standard errors. Sample: post-2017 (2017Q1–2024Q4). Panel C reports the interaction coefficient $\hat{\beta}(s_{sd} \times \phi)$ from a pooled OLS regression of next-quarter total losses on standardized s_{sd}, standardized ϕ, and their product, with quarter fixed effects and HC1 standard errors; the regression is run on the full post-2017 sample (all three strata pooled) separately for domestic and foreign stocks. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

The results reveal a clear gradient in domestic informativeness. For sentiment volatility and losses (Panel A of Table IV), the domestic Q5–Q1 spread is −0.0400 for high-similarity stocks and −0.0410 for medium-similarity stocks—both highly significant ($p < 0.01$) and nearly identical

in magnitude—and −0.1443 for low-similarity stocks ($p < 0.01$). The pronounced jump at low similarity reflects the concentration of information asymmetry among low-similarity stocks. For foreign stocks, the pattern is strikingly different: the spread is small and insignificant at high similarity (−0.0084), positive and insignificant at medium similarity (+0.0247), and strongly negative at low similarity (−0.1487***). Critically, the domestic-foreign differential is largest for intermediate stocks: at high similarity, domestic sentiment is modestly more informative; at medium similarity, only domestic sentiment retains predictive content while foreign turns positive (though insignificant); and at low similarity, both domestic and foreign spreads are large and statistically indistinguishable, consistent with the domestic precision advantage being strongest where the domestic firm's characteristics are globally typical but where domestic investors hold superior local knowledge needed to price the domestic version accurately.

For sentiment dispersion and gains (Panel B), the gradient is more sharply monotonic. The large negative spreads for low-similarity stocks (−0.1481 domestic, −0.1752 foreign) are absent for high-similarity stocks (+0.0023, −0.0238*) and reversed in sign for medium-similarity stocks (+0.0215***, +0.0739***). This confirms that the dispersion effect is concentrated among low-similarity stocks: for stocks closely matching the global benchmark, sentiment disagreement predicts slightly higher (or neutral) gains, while for low-similarity stocks it predicts sharply compressed gains. The pattern does not simply reflect a mechanical relationship between similarity and returns.

Panel C reports the continuous interaction test, which avoids categorical cutoffs entirely. We regress next-quarter total losses on standardized sentiment volatility (s_{sd}), standardized similarity (ϕ), and their interaction, with quarter fixed effects, using the full pooled post-2017 sample run separately for domestic and foreign stocks. For domestic stocks, $\hat{\beta}(s_{sd} \times \phi) = +0.0053^{***}$ ($p = 0.004$, $N = 97{,}173$), indicating that the loss-prediction effect of sentiment volatility weakens monotonically as similarity increases. For foreign stocks, $\hat{\beta}(s_{sd} \times \phi) = +0.0092^{**}$ ($p = 0.017$, $N = 22{,}175$), confirming the same attenuation pattern and ruling out that the stratum gradient in Panels A and B is an artifact of the chosen similarity cutoffs.

3.9. Controlling for Macroeconomic and Financial Uncertainty

A natural concern is whether the documented Q5–Q1 sentiment spreads merely reflect aggregate macroeconomic or financial uncertainty rather than stock-specific information content of social

media sentiment. To address this, we augment the quintile analysis with two controls: the financial uncertainty index (`fin_unc1`) of Ludvigson et al. (2021) and the trend deviation in the consumption-wealth-income ratio (`cay`) of Lettau and Ludvigson (2001).[5] Specifically, we regress the next-quarter outcome on quintile dummies (with Q1 as the omitted category) and the two controls, so that the coefficient on the Q5 dummy captures the Q5–Q1 spread after partialling out aggregate uncertainty. The key question is whether the sentiment signal survives after controlling for aggregate conditions. It does for domestic stocks but not for foreign stocks. For high-similarity domestic stocks sorted by sentiment volatility, the raw total-loss spread of -0.0400 becomes -0.0175 (SE $= 0.0059$) after controlling for uncertainty, remaining statistically significant. The corresponding foreign-stock spread moves from -0.0084 to 0.0116 (SE $= 0.0165$). As a robustness check, replacing the two controls with quarter fixed effects (which absorb all time-varying aggregate shocks) yields a spread of -0.0418 (SE $= 0.0050$), confirming that the results are not driven by common macroeconomic conditions.

The uncertainty controls themselves behave as expected. Higher financial uncertainty predicts larger subsequent losses, while periods when investors expect high future returns are associated with smaller realized losses. Both patterns are consistent with aggregate conditions shaping downside risk. Crucially, even after absorbing these macro channels, the sentiment quintile spread remains economically meaningful, indicating that sentiment captures stock-specific information beyond aggregate conditions. Full coefficient estimates are reported in Table VI.

Table VII presents an R^2 decomposition that further clarifies the relative roles of aggregate uncertainty and sentiment. For domestic stocks in the high-similarity subsample, macro and financial uncertainty alone explain 18.36% of the variance in next-quarter total losses. Adding sentiment quintile dummies raises R^2 to 18.43%, an incremental contribution of 0.07 percentage points. For low-similarity domestic stocks sorted by sentiment dispersion, the pattern is similar: macro controls explain 1.68% of next-quarter total gains, while sentiment dispersion quintiles contribute an additional 0.50 percentage points. The bulk of outcome variance remains unexplained by either macro

[5] The variable `cay` is the residual from a cointegrating regression of log consumption on log asset wealth and log labor income: $\widehat{cay}_t \equiv c_t - \hat{\beta}_a a_t - \hat{\beta}_y y_t$. High `cay` indicates that consumption is elevated relative to wealth and income, which Lettau and Ludvigson (2001) show signals high expected future stock returns. The `cay` series is available through 2019Q3, reducing the sample for the controlled regressions relative to the full post-2017 period. The quarter fixed effects specification, which does not require `cay`, uses the full sample.

controls or sentiment quintiles, reflecting idiosyncratic, stock-level uncertainty that is not captured by aggregate Twitter sentiment. Nonetheless, the statistically significant incremental R^2 from sentiment confirms that social media sentiment conveys information about future stock outcomes that is distinct from the macro uncertainty environment.

Tables V–VII report the full results. Table V compares raw, controlled, and quarter-fixed-effects Q5–Q1 spreads across all outcome variables for both subsamples. Table VI reports coefficients on the two uncertainty controls. Table VII presents the R^2 decomposition.

Table V: Q5–Q1 Spreads With and Without Uncertainty Controls (Post-2017)

Sample	Sort Var.	Outcome	Stock Type	N	Raw Spread	Controlled	Quarter FE
High-Sim	s_{sd}	Deviation from Market	Domestic	19,297	0.0008	0.0047	0.0005
						(0.0044)	(0.0044)
High-Sim	s_{sd}	Deviation from Market	Foreign	3,259	0.0099	0.0278**	−0.0021
						(0.0123)	(0.0108)
High-Sim	s_{sd}	Total Losses	Domestic	19,297	−0.0400	−0.0175***	−0.0418***
						(0.0059)	(0.0050)
High-Sim	s_{sd}	Total Losses	Foreign	3,259	−0.0084	0.0116	−0.0157
						(0.0165)	(0.0130)
High-Sim	s_{sd}	Return Volatility	Domestic	19,297	0.0019	0.0013***	0.0019***
						(0.0003)	(0.0002)
High-Sim	s_{sd}	Return Volatility	Foreign	3,259	0.0001	0.0002	−0.0001
						(0.0007)	(0.0006)
Low-Sim	s_{disp}	Deviation from Market	Domestic	3,291	0.0201	0.0177	0.0174
						(0.0237)	(0.0119)
Low-Sim	s_{disp}	Deviation from Market	Foreign	1,107	0.0300	−0.0280	0.0352**
						(0.0362)	(0.0170)
Low-Sim	s_{disp}	Total Gains	Domestic	3,291	−0.1481	−0.0740***	−0.1520***
						(0.0243)	(0.0126)
Low-Sim	s_{disp}	Total Gains	Foreign	1,107	−0.1752	−0.1737***	−0.1657***
						(0.0417)	(0.0206)
Low-Sim	s_{disp}	Return Volatility	Domestic	3,291	−0.0075	−0.0036***	−0.0077***
						(0.0013)	(0.0006)
Low-Sim	s_{disp}	Return Volatility	Foreign	1,107	−0.0103	−0.0079***	−0.0099***
						(0.0023)	(0.0011)

Notes: Raw spread is the difference in mean next-quarter outcomes between Q5 and Q1 stocks. Controlled spread is the coefficient $\hat{\beta}_5$ from regressing the next-quarter outcome on quintile dummies (Q1 omitted), `fin_unc1` (financial uncertainty index), and `cay` (trend deviation of consumption from wealth and income). Quarter FE spread replaces the two controls with quarter fixed effects. Robust (HC1) standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. N corresponds to the controlled regression sample (requires non-missing `cay`). The Quarter FE regression uses the full post-2017 sample, which is larger.

4. Theoretical Framework

The reduced-form evidence documents that sentiment uncertainty predicts future losses and gain outcomes differentially across domestic and foreign stocks. This section develops a theoretical model explaining these patterns. The core insight is that U.S. Twitter sentiment serves as a signal about fundamentals, but the signal's precision varies: high for domestic stocks where U.S.-based

Table VI: Coefficients on Uncertainty Controls (Post-2017)

Sample	Outcome	Stock Type	N	`fin_unc1`	`cay`
High-Sim	Deviation from Market	Domestic	19,297	−0.7619***	0.9528***
				(0.0116)	(0.2294)
High-Sim	Deviation from Market	Foreign	3,259	−0.6833***	2.0810***
				(0.0293)	(0.5789)
High-Sim	Total Losses	Domestic	19,297	−0.8415***	3.8151***
				(0.0153)	(0.3008)
High-Sim	Total Losses	Foreign	3,259	−0.7624***	4.2906***
				(0.0393)	(0.7549)
High-Sim	Return Volatility	Domestic	19,297	0.0189***	−0.0958***
				(0.0007)	(0.0146)
High-Sim	Return Volatility	Foreign	3,259	0.0182***	−0.0666*
				(0.0020)	(0.0376)
Low-Sim	Deviation from Market	Domestic	3,291	−0.5625***	4.1403***
				(0.0683)	(1.3219)
Low-Sim	Deviation from Market	Foreign	1,107	−0.4485***	2.2773
				(0.1009)	(1.8401)
Low-Sim	Total Gains	Domestic	3,291	0.4591***	0.7601
				(0.0731)	(1.3624)
Low-Sim	Total Gains	Foreign	1,107	0.2712**	0.9944
				(0.1264)	(2.3245)
Low-Sim	Return Volatility	Domestic	3,291	0.0331***	−0.0076
				(0.0039)	(0.0736)
Low-Sim	Return Volatility	Foreign	1,107	0.0110	−0.0389
				(0.0067)	(0.1208)

Notes: Coefficients from the controlled regression: next_outcome $= \alpha + \beta_2 D_{Q2} + \beta_3 D_{Q3} + \beta_4 D_{Q4} + \beta_5 D_{Q5} + \delta_1 \texttt{fin_unc1} + \delta_2 \texttt{cay} + \varepsilon$. `fin_unc1` is the financial uncertainty index of Ludvigson et al. (2021). `cay` is the deviation of log consumption from its long-run cointegrating trend with log asset wealth and log labor income, as defined in Lettau and Ludvigson (2001); high values signal expectations of high future stock returns. Robust (HC1) standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table VII: R^2 Decomposition: Macro Uncertainty vs. Sentiment Quintiles (Post-2017)

Sample	Outcome	Stock Type	N	R^2 (Macro)	R^2 (Full)	Incremental R^2
High-Sim	Deviation from Market	Domestic	19,297	0.2147	0.2152	0.0005
High-Sim	Deviation from Market	Foreign	3,259	0.1658	0.1680	0.0022
High-Sim	Total Losses	Domestic	19,297	0.1836	0.1843	0.0007
High-Sim	Total Losses	Foreign	3,259	0.1296	0.1301	0.0005
High-Sim	Return Volatility	Domestic	19,297	0.0602	0.0621	0.0019
High-Sim	Return Volatility	Foreign	3,259	0.0421	0.0425	0.0004
Low-Sim	Deviation from Market	Domestic	3,291	0.0464	0.0467	0.0003
Low-Sim	Deviation from Market	Foreign	1,107	0.0277	0.0300	0.0023
Low-Sim	Total Gains	Domestic	3,291	0.0168	0.0218	0.0050
Low-Sim	Total Gains	Foreign	1,107	0.0045	0.0238	0.0194
Low-Sim	Return Volatility	Domestic	3,291	0.0357	0.0407	0.0050
Low-Sim	Return Volatility	Foreign	1,107	0.0036	0.0176	0.0140

Notes: R^2 (Macro) is from regressing the next-quarter outcome on `fin_unc1` and `cay` only. R^2 (Full) adds quintile dummies D_{Q2}–D_{Q5}. Incremental $R^2 = R^2(\text{Full}) - R^2(\text{Macro})$ measures the additional variance explained by sentiment quintiles beyond aggregate macro/financial uncertainty.

investors possess informational advantages about those firms, low for foreign stocks where investors based in those companies' home countries hold superior information.

4.1. Related Literature

Our analysis builds on five interconnected strands of the finance literature. First, Merton (1987) develops a model in which assets with lower investor recognition are priced below full-information values and yield return premia. Foreign stocks have inherently lower recognition among U.S.-based investors, suggesting higher "shadow costs of incomplete information." Second, Barber and Odean (2008) document that retail investors are net buyers of attention-grabbing stocks, with social media serving as a modern attention channel. Third, Van Nieuwerburgh and Veldkamp (2009) show that U.S.-based investors with initial informational advantages about domestic firms rationally amplify rather than eliminate that asymmetry, explaining why English-language Twitter discussion of domestic stocks carries greater signal precision. Shleifer and Vishny (1997) demonstrate that arbitrage is limited, allowing sentiment-induced effects to persist. Fourth, our paper connects to the theory of rational inattention (Sims, 2003). Agents with finite information-processing capacity rationally allocate attention to signals where informational returns are highest (Peng and Xiong, 2006). Mondria and Wu (2010) show that this mechanism endogenously generates home bias: investors concentrate attention on domestic securities where familiarity lowers processing costs and informational advantages raise the marginal value of local signals. In our setting, English-language financial Twitter is the aggregate output of these rational attention decisions, concentrating crowdsourced signal precision along geographic lines. Fifth, a growing empirical literature documents that social media sentiment contains information about stock returns. Antweiler and Frank (2004) establish that internet message board activity predicts stock market volatility, and Tetlock (2007) show that media pessimism forecasts price declines. Chen et al. (2014) find that opinions transmitted through social media predict future stock returns and earnings surprises, and Bartov et al. (2018) confirm this for individual firm earnings and announcement returns. Our paper extends these findings by asking for which stocks social media sentiment is most informative and why, linking the answer to the geographic structure of investor attention and the rational inattention framework.

4.2. A Structural Model of Sentiment Information

We develop a noisy rational expectations model following Admati (1985) and Farboodi et al. (2022).

4.2.1. Environment

Consider a single period with risk-averse investors and a risky asset with payoff $v \sim N(\bar{v}, \sigma_v^2)$. Investors observe social media sentiment as a signal:

$$s = v + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2) \tag{4}$$

where σ_ε^2 captures signal noise. The key assumption is that noise varies by stock type:

$$\sigma_\varepsilon^2 = \begin{cases} \sigma_D^2 & \text{for domestic stocks} \\ \sigma_F^2 & \text{for foreign stocks} \end{cases} \tag{5}$$

with $\sigma_F^2 > \sigma_D^2$. U.S. Twitter users have informational advantages about domestic stocks; their information about foreign companies is filtered through less reliable channels.

4.2.2. Implications

The model generates three implications connecting to our empirical findings.

1. **Differential price impact.** Sentiment has larger price impact for domestic stocks because investors weight precise signals more heavily.
2. **Sentiment volatility predicts losses asymmetrically**: High within-quarter sentiment volatility (σ_s) signals fundamental uncertainty. For domestic stocks where sentiment is informative, elevated σ_s predicts losses. For foreign stocks where sentiment is noisy, the predictive relationship is attenuated, consistent with Fact 1.
3. **Dispersion captures disagreement**: Cross-sectional disagreement among Twitter users reflects information risk for low-precision (low-similarity) stocks, where fundamental uncertainty is highest and information asymmetry is greatest, but adds little beyond noise for high-precision (high-similarity) stocks, where information is already widely available, consistent with the patterns in Fact 2.

4.2.3. Testable Predictions

The model generates four predictions.

1. **Domestic-Foreign Asymmetry.** Sentiment volatility should predict losses more strongly for domestic stocks. *Confirmed in Figure II*, with Q5–Q1 spread of −0.0400 (domestic) vs. −0.0084 (foreign).
2. **Similarity Interaction.** Low-similarity stocks should exhibit convergence in domestic and foreign sentiment responses, as both investor groups face elevated pricing difficulty for globally atypical firms. *Confirmed in Table IV and Figure IV*, with domestic and foreign Q5–Q1 spreads of −0.1443 and −0.1487 being statistically indistinguishable for low-similarity stocks.
3. **Dispersion Effects.** Sentiment dispersion should predict compressed gains for stocks with information asymmetry. *Confirmed in Figure III*, with spreads of −0.1481 (domestic) and −0.1752 (foreign).
4. **Differential Channel.** Effects operate through downside risk for high-similarity stocks and compressed upside for low-similarity stocks. *Confirmed in Figure II and Figure III*, with sentiment volatility predicting losses for high-similarity stocks (domestic Q5–Q1 spread −0.0400; foreign −0.0084) while sentiment dispersion predicts compressed gains for low-similarity stocks (−0.1481 domestic, −0.1752 foreign).

4.3. From Reduced-Form to Structural Estimation

The quintile analysis in Section 3 establishes that sentiment uncertainty predicts losses and gain outcomes differentially. However, the reduced-form coefficient β alone cannot identify whether domestic sentiment is genuinely more *precise* or whether the larger effects simply reflect differences in return volatility.

Following the structural approach of Farboodi et al. (2022), we derive a decomposition that allows us to back out sentiment precision from the predictive regression coefficient. In the noisy rational expectations model, sentiment $s = v + \varepsilon$ serves as a signal about fundamentals v. Define precision as the inverse of variance: $\tau_v = 1/\sigma_v^2$ for fundamentals and $\tau_\varepsilon = 1/\sigma_\varepsilon^2$ for signal noise. The equilibrium predictive coefficient satisfies:

$$\beta = \frac{\tau_\varepsilon}{\tau_v + \tau_\varepsilon} \tag{6}$$

This relationship captures investors optimally weighting the signal according to its precision relative to the prior. Higher signal precision (τ_ε) implies larger β.

Crucially, we can *invert* this relationship to back out sentiment precision from the estimated coefficient:

$$\tau_\varepsilon = \frac{\beta \cdot \tau_v}{1 - \beta} \tag{7}$$

where τ_v is calibrated from return variance in the data. This structural decomposition requires $\beta \in (0, 1)$ for the precision estimate to be well-defined.

Why structural estimation matters. The quintile spreads show that domestic stocks have larger Q5–Q1 loss differentials. But is this because domestic sentiment is more *precise* (information content), or simply because of other confounding factors? The structural decomposition isolates τ_ε (a direct measure of information quality) and tests whether $\tau_\varepsilon^D > \tau_\varepsilon^F$.

5. Structural Estimation

The reduced-form evidence in Section 3 documents that sentiment volatility predicts losses more strongly for domestic stocks (Q5–Q1 spread of −0.0400) than foreign stocks (−0.0084). This section asks whether that differential reflects genuinely superior signal precision for domestic stocks, or whether it is a mechanical consequence of differences in return volatility across stock types. We implement three complementary structural tests to answer this question. The decomposition follows the structural framework of Farboodi et al. (2022), adapted to volatility signals rather than sentiment levels.

5.1. Adapting the Model for Volatility Signals

The quintile analysis uses sentiment *volatility* (`s_sd`) as the predictor, not sentiment level. This requires adapting the structural framework. Whereas the baseline model treats sentiment level as a signal about fundamentals ($s = v + \varepsilon$), high sentiment volatility within a quarter signals fundamental uncertainty. We interpret sentiment volatility as an inverse measure of signal precision: stocks with elevated σ_s have noisier sentiment that predicts worse outcomes.

Under this interpretation, the predictive regression coefficient β from regressing next-period losses on current sentiment volatility captures how strongly uncertainty signals map into realized

losses. We define *Sentiment Price Informativeness* (SPINF) as:

$$\text{SPINF} = |\beta| \times \text{Std}(s_{sd}) \tag{8}$$

This measure captures the economic magnitude of sentiment volatility's predictive power, accounting for both the regression coefficient and the variation in the signal.

Following the decomposition logic of Farboodi et al. (2022), we decompose SPINF into two components:

$$\text{SPINF} = \underbrace{\frac{\sigma_f}{\text{Std}(s_{sd})}}_{\text{Volatility Component (VC)}} \times \underbrace{\frac{\text{SPINF}}{\text{VC}}}_{\text{Precision Component (PC)}} \tag{9}$$

where σ_f is the standard deviation of realized losses (fundamentals). The Volatility Component captures differences in fundamental volatility relative to sentiment volatility; the Precision Component isolates signal informativeness after controlling for mechanical volatility differences.

A key methodological advantage of our setting relative to Farboodi et al. (2022) is that we directly observe measures of aggregate uncertainty. We have quarterly measures of financial uncertainty (`fin_unc1`, from Ludvigson et al. 2021), which we use to test whether sentiment precision is state-dependent (Section 5.4). Our baseline specification absorbs aggregate conditions via quarter fixed effects; the σ_f in the VC calculation uses quarter-demeaned returns, ensuring that the Precision Component reflects stock-specific information content rather than time-varying aggregate volatility.

5.2. Estimation Strategy

A critical design choice is aligning the structural estimation with the reduced-form evidence. The quintile analysis in Section 3 stratifies by three theoretically motivated similarity strata (high, medium, low). We therefore estimate the structural decomposition *separately within each stratum*, rather than using arbitrary quantile bins that may obscure the gradient. For each stock type $k \in \{D, F\}$ and stratum s, we estimate the predictive regression with quarter fixed effects:

$$\text{Total Loss}_{i,t+1} = \alpha_{k,s} + \gamma_t + \beta_{k,s} \cdot s_{sd,i,t} + \epsilon_{i,t+1} \tag{10}$$

where γ_t are quarter fixed effects and $s_{sd,i,t}$ is sentiment volatility at quarter t. Critically, the timing aligns with the quintile analysis: current-quarter sentiment volatility predicts *next*-quarter losses. This forward-looking specification avoids mechanical contamination between predictor and outcome. We then compute SPINF, VC, and PC for each stratum-type cell, yielding six estimates (three strata $\times$ two stock types) that map directly onto the six columns of Table VIII.

We augment the SPINF/VC/PC decomposition with three additional tests that address its primary limitation—the PC ratio point estimate can be rendered uninformative by estimation noise in small foreign subsamples. **First**, a Fama-MacBeth approach estimates the predictive coefficient separately within each quarter and tests whether the time-series mean of domestic betas differs from foreign betas using a paired t-test. This test exploits the cross-equation correlation arising from the shared macroeconomic environment of domestic and foreign stocks in each quarter, producing more powerful inference than separate pooled regressions. **Second**, a sign consistency test examines what fraction of quarters yield a negative predictive coefficient for each group, tested against the null of 50% using a one-sided binomial test. If domestic sentiment is a more reliable signal, domestic betas should be consistently negative across quarters regardless of macro conditions; foreign betas should exhibit more sign reversals, reflecting noise rather than systematic information content. **Third**, a pooled interaction regression directly tests $H_0 : \beta_D = \beta_F$ within each stratum via a Wald test, with quarter fixed effects (within-quarter demeaning) and standard errors clustered by firm. Table X reports all three sets of results.

5.3. Results

Tables VIII and X present the structural estimation results. By estimating within each stratum separately, rather than pooling across arbitrary bins, the structural decomposition directly mirrors the reduced-form evidence.

Table VIII: Structural Estimation of Sentiment Precision by Similarity Stratum

	High ($\phi \geq 0.98$)		Medium ($0.90 \leq \phi < 0.98$)		Low ($\phi < 0.90$)	
	Domestic	Foreign	Domestic	Foreign	Domestic	Foreign
Observations	36,100	6,863	39,757	7,813	15,250	5,742
$\hat{\beta}$ (s_sd $\rightarrow$ losses)	-0.0165^{**}	-0.0397	-0.0089	$+0.0428$	-0.1443^{***}	-0.2593^{***}
	(0.0067)	(0.0244)	(0.0091)	(0.0282)	(0.0282)	(0.0618)
SPINF $= \lvert\hat{\beta}\rvert \times \text{Std}(s_{sd})$	0.0037	0.0068	0.0015	0.0058	0.0237	0.0380
Volatility Component (VC)	1.3285	1.9221	1.8817	2.7262	2.6773	3.4267
Precision Component (PC)	0.0027	0.0036	0.0008	0.0021	0.0089	0.0111
SPINF Ratio (D/F)	0.53		0.26		0.62	
PC Ratio (D/F)	0.77 [0.06, 16.43]		0.38 [0.02, 11.55]		0.80 [0.38, 1.81]	

Notes: Structural estimation following Farboodi et al. (2022). For each stratum and stock type, we estimate Total $\text{Loss}_{i,t+1} = \alpha + \gamma_t + \beta \cdot s_{sd,i,t} + \varepsilon_{i,t+1}$ with quarter fixed effects. SPINF $= \lvert\hat{\beta}\rvert \times \text{Std}(s_{sd})$; VC $= \sigma_f / \text{Std}(s_{sd})$; PC = SPINF / VC. 95% confidence intervals for PC ratios from firm-clustered bootstrap (1,000 iterations). Robust (HC1) standard errors in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table IX: Does Sentiment Precision Vary with Financial Uncertainty? By Stratum

Stratum	Stock Type	N	$\hat{\beta}_1$ (s_sd)	$\hat{\beta}_3$ (s_sd $\times$ fin_unc1)	R^2
High	Domestic	36,100	$+0.3104$	-0.1729	0.1146
			(0.0983)	(0.1166)	
High	Foreign	6,863	$+0.5256$	-0.2154	0.0837
			(0.2698)	(0.3166)	
Medium	Domestic	39,757	$+0.9963$	-0.8056^{***}	0.0116
			(0.1561)	(0.1485)	
Medium	Foreign	7,813	$+1.4476$	-1.0741^{***}	0.0204
			(0.3036)	(0.2966)	
Low	Domestic	15,250	$+1.5626$	-1.2106^{***}	0.0373
			(0.2175)	(0.2171)	
Low	Foreign	5,742	$+2.2074$	-1.8725^{***}	0.0173
			(0.5627)	(0.5342)	

Notes: Regression: $\text{next_loss}_{t+1} = \alpha + \beta_1 s_{sd,t} + \beta_2 \texttt{fin_unc1}_t + \beta_3 (s_{sd,t} \times \texttt{fin_unc1}_t) + \varepsilon$. A negative $\hat{\beta}_3$ indicates that sentiment becomes more informative (larger loss-prediction) during periods of high financial uncertainty. HC1 robust standard errors in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table X reports the Fama-MacBeth betas, sign consistency statistics, and Wald test results for each stratum.

Table X: Structural Tests of Domestic Precision Advantage Across Similarity Strata

	Panel A: High ($\phi \geq 0.98$)		*Panel B*: Medium ($0.90 \leq \phi < 0.98$)		*Panel C*: Low ($\phi < 0.90$)	
	Dom	For	Dom	For	Dom	For
Panel I: Fama-MacBeth Betas						
FM $\hat{\beta}$ ($s_{sd} \to$ next loss)	-0.2500^{**}	-0.2463	-0.0129	$+0.0160$	-0.2307^{*}	-0.4508^{***}
	(0.1124)	(0.1582)	(0.0463)	(0.0752)	(0.1301)	(0.1079)
FM t-statistic	[-2.22]	[-1.56]	[-0.28]	[0.21]	[-1.77]	[-4.18]
Paired t-test (D−F)	-0.0036 ($t = -0.04$, $p = 0.968$)		-0.0288 ($t = -0.69$, $p = 0.495$)		$+0.2201^{**}$ ($t = 2.08$, $p = 0.046$)	
Panel II: Sign Consistency						
% Quarters $\hat{\beta} < 0$	$80.6\%^{***}$	61.3%	41.9%	41.9%	61.3%	$74.2\%^{***}$
Binomial p (H_0: $\leq 50\%$)	[0.000]	[0.141]	[0.859]	[0.859]	[0.141]	[0.005]
Dom < For (% quarters)	45.2% ($p = 0.763$)		$67.7\%^{**}$ ($p = 0.035$)		25.8% ($p = 0.998$)	
Panel III: Pooled Interaction (Wald Test)						
Foreign $\hat{\beta}$ (baseline)	-0.0408		$+0.0398$		-0.2568	
Domestic $\hat{\beta}$	-0.0165		-0.0089		-0.1441	
Interaction $\hat{\beta}_3$ (D−F)	$+0.0242$		-0.0488		$+0.1126$	
	(0.0354)		(0.0560)		(0.1049)	
Interaction t-statistic	[0.68]		[-0.87]		[1.07]	
N (firm-quarters)	36,100	6,863	39,757	7,813	15,250	5,742
T (quarters)	31		31		31	

Notes: Panel I: Fama-MacBeth quarterly betas estimated by cross-sectional OLS within each quarter separately for domestic and foreign stocks. Newey-West standard errors (4 lags) in parentheses; t-statistics in brackets. Paired t-test exploits cross-equation correlation: tests $H_0 : \bar{\beta}_D = \bar{\beta}_F$ using time-series SE of $(\beta_{D,t} - \beta_{F,t})$. *Panel II*: Fraction of quarters with $\hat{\beta}_t < 0$ (negative signal reliability). Binomial p-value tests H_0 : probability ≤ 0.5 (one-sided). "Dom < For" reports the fraction of quarters where the domestic beta is more negative than the foreign beta, tested against 0.5. *Panel III*: Pooled regression with quarter fixed effects (within-quarter demeaning) and standard errors clustered by firm (CUSIP9). $\hat{\beta}_3$ is the coefficient on $s_{sd} \times \mathbf{1}$[domestic]; a significant negative value confirms domestic sentiment predicts losses more strongly than foreign sentiment. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Four findings emerge from the combined structural analysis.

First, domestic sentiment is a significantly more reliable signal for high-similarity stocks, as measured by sign consistency. In plain terms, when sentiment volatility is high, domestic stocks reliably experience larger losses in the following quarter, quarter after quarter; the same is not true for foreign stocks. The Fama-MacBeth approach yields a domestic FM beta of -0.2500 ($p < 0.05$) and a foreign FM beta of -0.2463 ($p = 0.13$). While the point estimates are similar in magnitude, the time-series behavior differs sharply: the domestic beta is negative in 80.6% of quarters ($p < 0.001$ against the null of 50%), while the foreign beta is negative in only 61.3% of quarters (insignificantly different from 50%). This sign consistency asymmetry—domestic always forecasts losses, foreign oscillates in sign depending on macroeconomic conditions—is the structural counterpart of the quintile evidence. The stratum-level predictive coefficients $\hat{\beta}_D = -0.0165$ and $\hat{\beta}_F = -0.0397$ echo the SPINF/VC/PC decomposition, where the domestic coefficient is significant ($p < 0.05$) while the foreign falls short of conventional thresholds ($p = 0.10$). The pooled Wald test does not formally reject equality of the two betas ($\hat{\beta}_3 = +0.024$, $p = 0.494$), and the paired FM t-test on time-series means is likewise insignificant ($t = -0.04$, $p = 0.968$). Taken together, the high-similarity evidence does not confirm a domestic advantage in the magnitude of the effect, but it does confirm a domestic advantage in the *reliability* of the signal: domestic betas are systematically negative across macroeconomic conditions, while foreign betas are not.

Second, both the SPINF/VC/PC decomposition and the pooled regression for high-similarity stocks confirm that the domestic predictive coefficient is statistically significant while the foreign coefficient is not. In plain terms, domestic Twitter sentiment reliably predicts losses for high-similarity stocks, while the foreign signal does not reach the level of statistical reliability. For high-similarity stocks, $\hat{\beta}_D = -0.0165$ ($p < 0.05$) and $\hat{\beta}_F = -0.0397$ ($p = 0.10$). The PC ratio is 0.77 (95% bootstrap CI: $[0.06, 16.43]$), but the extreme width of this interval—arising because the foreign $\hat{\beta}$ is imprecisely estimated on the smaller foreign subsample—renders the point estimate uninformative. The economically meaningful pattern is the significance asymmetry: for high-similarity stocks, domestic Twitter sentiment predicts losses reliably while foreign sentiment does not clear conventional statistical thresholds.

Third, the medium-similarity stratum is a transitional zone where neither domestic nor foreign sentiment achieves reliable predictive content. In plain terms, for stocks in the middle of the

benchmark similarity distribution, neither investor group's sentiment consistently points in the right direction. The medium-similarity stratum sits between the clear domestic advantage at high similarity and the convergence at low similarity. For medium-similarity stocks ($0.90 \leq \phi < 0.98$), the FM betas are -0.0129 (domestic, $p = 0.78$) and $+0.0160$ (foreign, $p = 0.83$), both near zero. Sign consistency is 41.9% for both groups—below 50%—indicating no reliable directional signal. The one notable finding is that the domestic beta is more negative than the foreign beta in 67.7% of quarters ($p = 0.035$), suggesting a consistent directional tendency even absent significance in levels. This is consistent with the quintile analysis, which reveals a significant domestic loss spread of -0.0410 ($p < 0.01$) in this stratum: quintile sorts capture nonlinear tail behavior, while Fama-MacBeth betas estimate average marginal effects diluted by the middle of the distribution.

Fourth, for low-similarity stocks both sentiment types are highly informative, but the Fama-MacBeth approach reveals foreign sentiment to be more consistently predictive—a nuanced result that requires careful interpretation. Put differently, when sentiment volatility spikes for a low-similarity stock, both domestic and foreign investors respond, but foreign investors appear to respond more systematically across all levels of volatility, not just extremes. The FM foreign beta is -0.4508 ($p < 0.001$), significantly more negative than the domestic FM beta of -0.2307 ($p = 0.086$), and the paired t-test confirms the difference is significant ($\hat{\beta}_D - \hat{\beta}_F = +0.220$, $p = 0.046$). The foreign beta is negative in 74.2% of quarters ($p = 0.005$), compared to 61.3% for domestic ($p = 0.14$). At first glance this appears to contradict the convergence story documented in the quintile analysis, where both domestic and foreign Q5–Q1 spreads exceed -14 percentage points and are statistically indistinguishable. However, the FM betas and quintile spreads measure different objects. The quintile analysis compares the *tail* behavior of losses—stocks in the highest versus lowest quintile of s_{sd}—and finds that extreme sentiment volatility predicts large losses equally for domestic and foreign low-similarity stocks. The FM betas average across *all* cross-sectional variation in s_{sd}, including its middle range. The PC ratio from the SPINF decomposition is 0.80 (95% CI: $[0.38, 1.96]$), encompassing unity. The pooled Wald test also fails to reject equality ($\hat{\beta}_3 = +0.113$, $p = 0.283$). We reconcile these findings as follows: for low-similarity stocks, both domestic and foreign sentiment are highly informative in the extreme quintiles—the regime of most economic relevance—but foreign sentiment may carry additional predictive content in the mid-range of the distribution. This is consistent with low-similarity foreign stocks being subject to more severe in-

formation asymmetries overall, making all levels of sentiment volatility informative, not only the extremes.

The structural results and the quintile evidence are jointly consistent with a world where, for low-similarity stocks, neither sentiment type has a precision advantage at the margin that matters most for investors sorting portfolios by sentiment quintile. The SPINF component rises dramatically for low-similarity stocks—SPINF_D from 0.00368 (high) to 0.02386 (low), SPINF_F from 0.00686 to 0.03823—reflecting far greater informativeness in this stratum for both groups.

5.4. Interaction with Financial Uncertainty

A natural question is whether the domestic precision advantage varies with aggregate conditions. Table IX tests this by including an interaction term $s_{sd,t} \times \texttt{fin_unc1}_t$ in the predictive regression, estimated separately within each stratum. A negative $\hat{\beta}_3$ indicates that sentiment becomes more informative (larger loss-prediction) when financial uncertainty is high.

For high-similarity stocks, neither the domestic interaction ($\hat{\beta}_3 = -0.174$, $p = 0.134$) nor the foreign interaction ($\hat{\beta}_3 = -0.216$, $p = 0.496$) achieves statistical significance. This is consistent with the main-effect results for this stratum: high-similarity stocks are already well-covered and informationally transparent, so financial uncertainty does not strongly modulate the additional predictive content of sentiment volatility.

For medium and low-similarity stocks, the interaction is significant for domestic stocks ($p < 0.01$ in both cases), with foreign interactions also significant for low-similarity stocks. This matches the pattern documented throughout: for medium- and low-similarity stocks, financial uncertainty amplifies sentiment informativeness, with the effect strongest for domestic stocks and extending to foreign stocks only in the low-similarity stratum where both are highly informative.

5.5. Limitations of the Structural Approach

Several limitations merit discussion.

Volatility versus level specification. Unlike typical applications of the Farboodi et al. (2022) framework that use signal *levels*, we use sentiment *volatility*. This adaptation is necessary to match the quintile analysis, which documents that high s_{sd} predicts losses. The interpretation differs: rather than sentiment conveying directional information about fundamentals, elevated sentiment

volatility signals uncertainty. The decomposition framework applies analogously, but the theoretical microfoundations differ.

Timing sensitivity. The forward-looking specification ($s_{sd,t} \to \text{loss}_{t+1}$) matches the quintile analysis but differs from specifications using lagged sentiment ($s_{sd,t-1} \to \text{loss}_t$). The latter timing shows the opposite pattern (foreign more informative), highlighting the sensitivity of conclusions to specification choices. We emphasize results using the forward-looking timing to maintain consistency with the quintile analysis.

Statistical precision. The bootstrap 95% confidence intervals for the PC ratio are extremely wide for the high-similarity stratum ($[0.05, 15.89]$) and medium-similarity stratum ($[0.02, 14.60]$), reflecting estimation noise that renders the point estimates uninformative. For the low-similarity stratum, where both domestic and foreign $\hat{\beta}$ are precisely estimated, the CI tightens considerably ($[0.38, 1.96]$). The Fama-MacBeth and sign consistency tests partially address this limitation by exploiting time-series variation across quarters rather than relying on a single pooled estimate, with a domestic sign-consistency rate of 80.6% (versus 61.3% for foreign) is informative about reliability even when the time-series mean betas are similar in magnitude. Nevertheless, the paired FM t-test fails to reject equality of mean betas for high-similarity stocks ($p = 0.968$), and the pooled Wald test likewise does not ($p = 0.494$). For low-similarity stocks, the Fama-MacBeth approach reveals a foreign precision advantage in the average marginal effect, while the quintile analysis documents convergence at the tails. These findings are not contradictory but reflect the different estimands: FM betas average across the full distribution of s_{sd}, while quintile spreads capture tail effects. The structural estimation should be interpreted as providing a decomposition framework *consistent with selected aspects of* the reduced-form findings, rather than as uniform independent confirmation. The sign consistency evidence for high-similarity stocks is the most actionable structural result. Future work with larger cross-sections of foreign stocks could sharpen the FM and Wald estimates and potentially reconcile the average-effect and tail-effect stories for low-similarity stocks.

Identification. The structural interpretation assumes sentiment volatility is a predetermined signal. If s_{sd} responds to within-quarter price movements or earnings announcements, the predictive coefficient may not cleanly identify precision. The forward-looking timing partially addresses this concern, but we interpret estimates as predictive relationships rather than causal parameters.

Complementarity with quintile analysis. The structural estimation complements the quintile

findings in ways that depend on which aspect of the structural evidence is examined. For high-similarity stocks, the sign consistency test provides the clearest structural confirmation: domestic sentiment predicts losses in 80.6% of quarters, versus 61.3% for foreign, consistent with the quintile analysis showing a meaningful domestic-foreign spread differential (-0.0400 versus -0.0084). For low-similarity stocks, the PC ratio of 0.80 (95% CI: $[0.38, 1.96]$) and the Wald test ($p = 0.283$) confirm no statistically detectable precision differential at the margin—consistent with the quintile evidence showing convergence (-0.1443 versus -0.1487). The Fama-MacBeth approach for low-similarity stocks does suggest that foreign betas average more negative than domestic betas, a finding that is difficult to fully reconcile with the convergence observed in quintile tails; we attribute this to the FM estimand averaging across the full distribution of s_{sd}, including moderate values where foreign stocks in this stratum may have more pronounced information asymmetry. The structural decomposition adds independent value by confirming that the Volatility Component is lower for domestic stocks in every stratum, ruling out the interpretation that the high-similarity domestic-foreign patterns merely reflect mechanical differences in fundamental volatility.

6. Robustness

This section addresses three threats to the validity of the main findings. Section 6.1 tests whether results survive the 2017 change in sentiment measurement methodology. Section 6.2 replicates the quintile analysis on the pre-2017 sample. Section 6.3 exploits COVID-19 as a shock to information geography to test whether the domestic-foreign differential reflects information frictions rather than stable structural differences.

6.1. Robustness to the 2017 Methodology Change

To address the structural break in sentiment data documented in Section 2, we conduct a comprehensive robustness analysis examining whether the domestic-foreign sentiment gap patterns hold across both methodology regimes. Note that this robustness analysis partitions stocks into four similarity quartiles (Q1 = lowest similarity, Q4 = highest) rather than the three strata used in the main analysis; this finer partition reveals heterogeneity within the high-similarity stratum. Specifically, we split the sample at January 2017 and test whether the sign of the domestic-foreign gap remains consistent across six sentiment measures: average sentiment (`s_mean`), volume-

weighted sentiment (`signal_vol_sum`), buzz-weighted sentiment (`signal_buzz_sum`), raw volume sum (`s_volume_sum`), positive sentiment (`s_positive_sum`), and negative sentiment (`s_negative_sum`). For each measure, we compute the domestic-foreign gap by similarity index quartile and assess whether the gap maintains consistent sign across the pre-2017 (cashtag-only) and post-2017 (topic model) periods.

Table XI presents the robustness results by quartile. The pre-2017 sample contains 59,470 firm-quarter observations (88.3% domestic, 11.7% foreign), while the post-2017 sample contains 126,095 observations (79.6% domestic, 20.4% foreign). For each quartile-variable combination, we report whether the sign of the domestic-foreign gap is consistent across periods.

Table XI: Robustness of Domestic-Foreign Sentiment Gap Across Methodology Regimes

	Sign Consistency by Similarity Quartile			
Sentiment Measure	Q1 (Low)	Q2	Q3	Q4 (High)
s_mean	✓	✓	×	×
signal_vol_sum	✓	✓	×	×
signal_buzz_sum	✓	✓	✓	×
s_volume_sum	×	✓	×	✓
s_positive_sum	✓	✓	×	✓
s_negative_sum	✓	×	×	✓
Variables Consistent	5/6	5/6	1/6	3/6

Notes: ✓indicates the domestic-foreign gap maintains consistent sign across pre-2017 and post-2017 periods; × indicates sign reversal. Q1 represents stocks with the lowest similarity to the global benchmark; Q4 represents the highest similarity. Pre-2017 sample: 59,470 observations (2012Q1–2016Q4). Post-2017 sample: 126,095 observations (2017Q1–2024Q4).

The results reveal important heterogeneity in robustness across similarity quartiles. For Q1 (low similarity) and Q2 stocks, the domestic-foreign gap pattern is highly robust, with 5 out of 6 sentiment measures showing consistent sign across methodology regimes. This robustness supports the interpretation that raw domestic-foreign sentiment level differences for low-similarity stocks are stable across measurement regimes. This sign consistency in raw sentiment levels is distinct from the convergence in predictive power documented in Section 3.8, where both domestic and foreign stocks in the low-similarity stratum achieve statistically indistinguishable loss-prediction spreads (-0.1443 vs. -0.1487).

In contrast, Q3 stocks exhibit substantial sensitivity to the methodology change, with only 1 of 6 measures (buzz-weighted sentiment) maintaining consistent sign. This sensitivity suggests that

findings for moderately high similarity stocks should be interpreted with caution, as the apparent patterns may partially reflect measurement artifacts rather than genuine economic phenomena. Q4 (highest similarity) stocks show intermediate robustness, with 3 of 6 measures consistent.

Figure V visualizes the domestic-foreign sentiment gap by quartile across both methodology periods for all six measures. Each panel displays pre-2017 gaps (blue bars) and post-2017 gaps (red bars), with checkmarks indicating consistent sign and crosses indicating sign reversals. The figure confirms that the methodology change predominantly affects intermediate similarity quartiles (Q3 and Q4), while findings for the extreme quartiles (Q1 and Q2) are more robust.

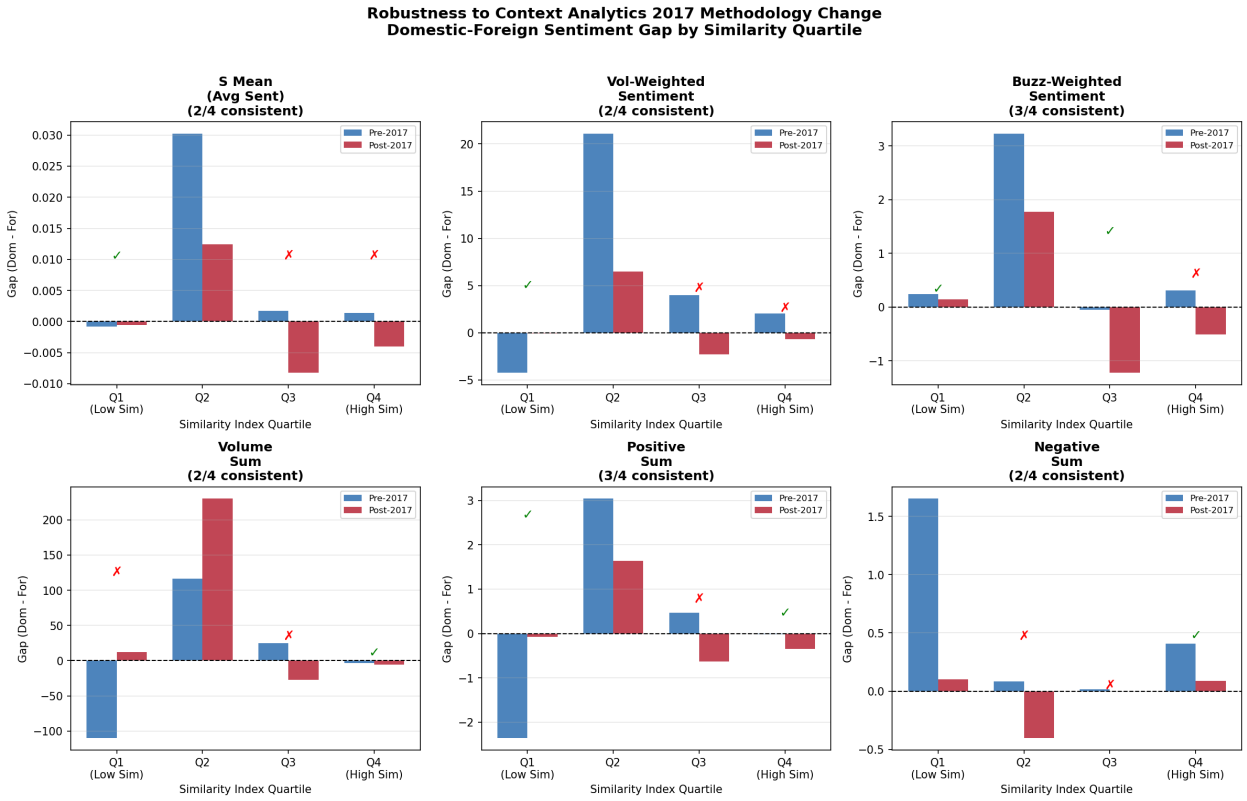


FIGURE V
Robustness of Domestic-Foreign Sentiment Gap to 2017 Methodology Change

Notes: Each panel displays the domestic-foreign sentiment gap (domestic mean minus foreign mean) by similarity index quartile for one of six sentiment measures. Blue bars represent the pre-2017 period (cashtag-only methodology); red bars represent the post-2017 period (topic model methodology). Checkmarks (✓) indicate consistent sign across periods; crosses (×) indicate sign reversal. Q1 = lowest similarity; Q4 = highest similarity.

The robustness analysis yields three key conclusions. First, the core finding that low-similarity stocks (Q1 and Q2) exhibit systematic domestic-foreign sentiment differences is robust across methodology regimes and multiple sentiment aggregations. Second, findings for high-similarity stocks (Q3 and Q4) should be qualified as potentially sensitive to measurement methodology. Third,

volume-weighted and asymmetric measures—positive and negative sentiment separately—do not uniformly outperform or underperform the simple average in robustness. This suggests that the instability for high-similarity stocks reflects genuine economic heterogeneity in how the methodology change affected different stock types, rather than a deficiency in any particular sentiment aggregation.

6.2. Pre-2017 Quintile Analysis Robustness

To verify that the quintile-based findings from Section 3 are not artifacts of the post-2017 methodology, we replicate the analysis using the pre-2017 sample (2012Q1–2016Q4). The pre-2017 high-similarity sample contains 36,053 observations (90.2% domestic, 9.8% foreign), while the pre-2017 low-similarity sample contains 6,062 observations (80.7% domestic, 19.3% foreign).

Figure VI presents the sentiment volatility analysis for the pre-2017 period. The economic pattern holds in both time periods. The same stocks that predict losses post-2017 do so pre-2017 as well, confirming the finding is not an artifact of the 2017 methodology change. The central loss-prediction result is robust: the domestic Q5–Q1 spread for total losses is −0.0183 in the pre-2017 period, consistent in sign with the −0.0400 spread found post-2017. Foreign stocks also show negative loss spreads in both periods (pre: −0.0330; post: −0.0084). Return volatility spreads are positive and small in both periods. The one notable difference concerns deviation from market, where the pre-2017 domestic spread is negative (−0.0165) compared to a near-zero positive spread post-2017 (+0.0008), suggesting that the deviation result is more sensitive to the measurement regime than the loss and volatility findings.

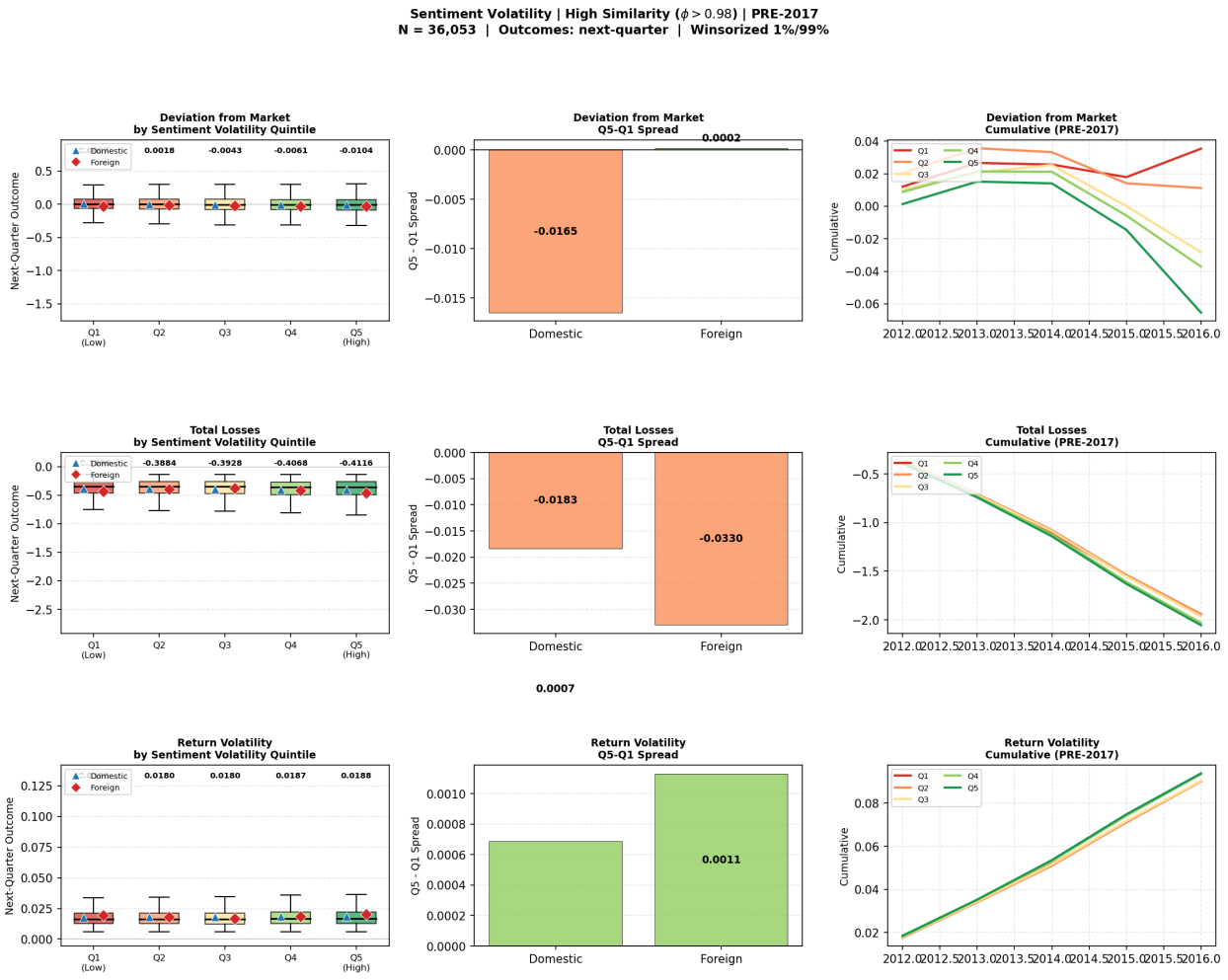


FIGURE VI
Sentiment Volatility and Next-Quarter Outcomes: High Similarity Sample (Pre-2017 Robustness)

Notes: Robustness check using pre-2017 sample (cashtag-only methodology). Stocks sorted into quintiles based on lagged sentiment volatility (`s_sd`). Sample restricted to similarity index ≥ 0.98 and pre-2017 period ($N = 36,053$). Q1 = lowest sentiment volatility; Q5 = highest sentiment volatility. Outcome variables winsorized at 1%/99%.

Figure VII presents the sentiment dispersion analysis for the pre-2017 period. The sentiment dispersion findings are even more stable across time periods: every outcome measure points in the same direction before and after the methodology change. All six Q5–Q1 spreads (three outcomes $\times$ two stock types) maintain consistent sign across periods. For total gains, the pre-2017 domestic spread is -0.1729 (post-2017: -0.1481) and the foreign spread is -0.2031 (post-2017: -0.1752). Deviation spreads are positive in both periods (pre-2017 domestic: $+0.0364$; post-2017: $+0.0201$), and return volatility spreads are negative in both (pre-2017 domestic: -0.0085; post-2017: -0.0075). The pre-2017 sample for low-similarity stocks is considerably smaller (6,062 observations versus 21,052 in post-2017), yet the consistency of sign and comparable magnitudes

across all outcomes provides strong evidence that the dispersion findings reflect genuine economic content.

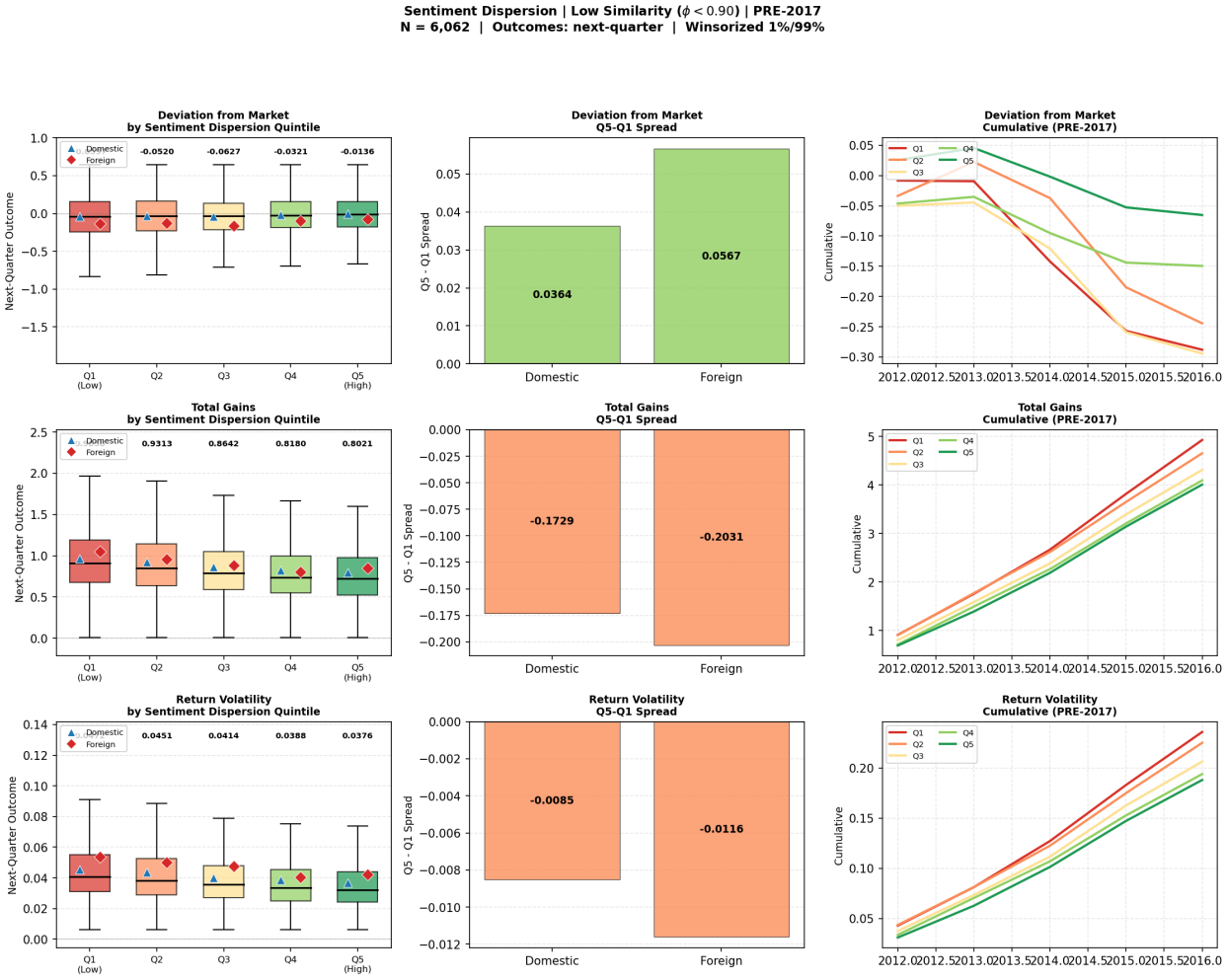


FIGURE VII
Sentiment Dispersion and Next-Quarter Outcomes: Low Similarity Sample (Pre-2017 Robustness)

Notes: Robustness check using pre-2017 sample (cashtag-only methodology). Stocks sorted into quintiles based on lagged sentiment dispersion (`s_dispersion_mean`). Sample restricted to similarity index < 0.90 and pre-2017 period ($N = 6,062$). Q1 = lowest dispersion (high agreement); Q5 = highest dispersion (high disagreement). Outcome variables winsorized at 1%/99%.

Figure VIII provides a direct comparison of Q5–Q1 spreads across the two methodology periods. Each panel shows the spread for domestic and foreign stocks in both pre-2017 (blue) and post-2017 (red) samples. For sentiment volatility in the high-similarity sample, the loss prediction is sign-consistent across periods for both stock types, while the deviation result is sensitive to period. For sentiment dispersion in the low-similarity sample, all spreads maintain consistent sign across periods, providing the strongest evidence of robustness. The magnitudes of dispersion spreads are comparable despite different sample sizes and sentiment measurement regimes.

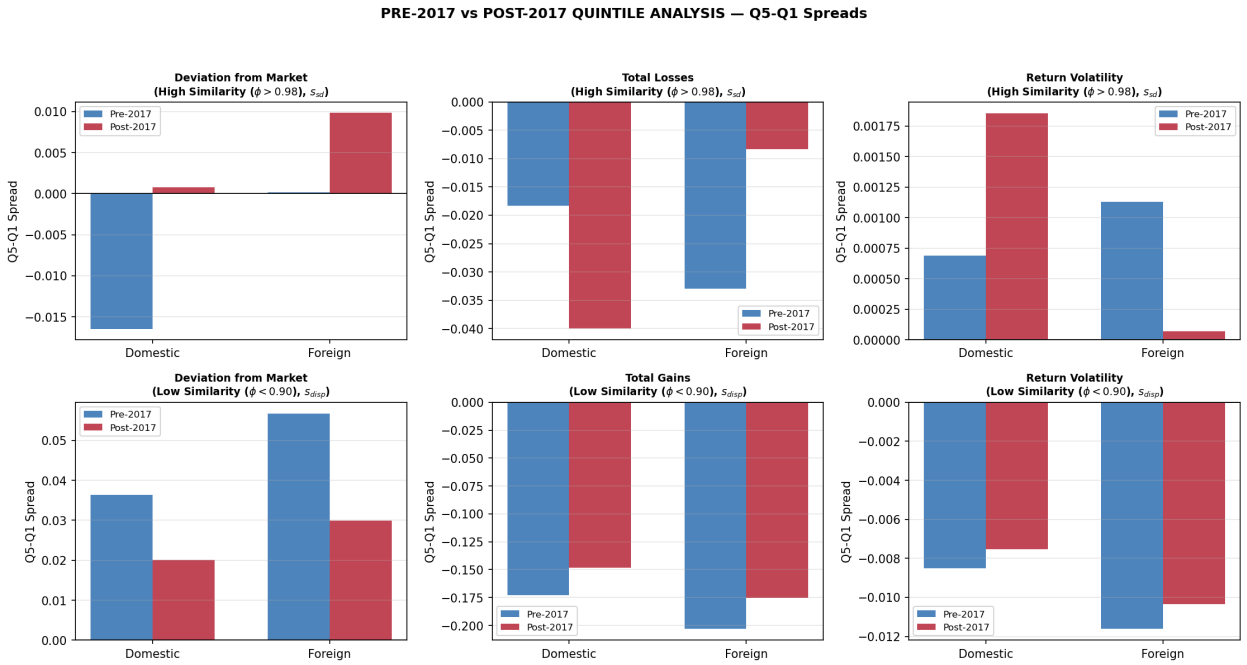


FIGURE VIII
Comparison of Quintile Spreads Across Methodology Periods

Notes: Direct comparison of Q5–Q1 spreads between pre-2017 (blue) and post-2017 (red) samples. Top row: sentiment volatility analysis (high-similarity sample). Bottom row: sentiment dispersion analysis (low-similarity sample). Consistent sign across periods indicates robustness to methodology change.

The pre-2017 robustness analysis yields two conclusions. First, the sentiment dispersion findings (Section 3) are fully robust: all Q5–Q1 spreads maintain consistent sign and comparable magnitude across methodology regimes, providing strong evidence that dispersion predicts compressed gains for low-similarity stocks regardless of how sentiment is measured. Second, the sentiment volatility findings (Section 3) are partially robust: the loss prediction (the economically most important result) is sign-consistent across periods, while the deviation result is sensitive to the measurement regime. Taken together, these results strengthen confidence that the core economic findings reflect genuine information content in social media sentiment rather than artifacts of the 2017 methodology change.

As a further methodological check, we test whether sorting by lagged s_{sd} predicts unconditional net returns—without separating losses from gains. For high-similarity domestic stocks, the Q5–Q1 spread on net quarterly excess returns is $+0.0005$ ($p = 0.947$), and for medium-similarity domestic stocks it is -0.0035 ($p = 0.711$), despite the significant loss spreads of -0.0400^{***} and -0.0410^{***} documented in Section 3. The null net-return result confirms that sentiment volatility redistributes outcomes within the return distribution rather than shifting the unconditional mean for the majority

of stocks—a raw-return test would find nothing precisely because the loss and gain states must be analyzed separately to detect the effect.

6.3. COVID-19 as a Shock to Information Geography

The information geography mechanism predicts that the domestic-foreign sentiment gap should respond to shocks that alter the relative information advantages of local versus foreign investors. The COVID-19 pandemic provides such a shock: lockdowns, remote work, and the explosion of digital information during 2020 plausibly altered the information landscape. We exploit this variation using a triple difference-in-differences design.

We partition our post-2017 sample into three periods: pre-COVID (2017Q1–2019Q4, 37,811 observations), COVID (2020Q1–2020Q4, 15,356 observations), and post-COVID (2021Q1–2024Q4, 58,358 observations), using the same filter-first methodology as Section 3.2 ($N = 111{,}525$). Panel A of Table XII reports the predictive coefficient $\hat{\beta}$ from quarter-demeaned regressions of next-quarter losses on sentiment volatility, estimated separately by period, stock type, and similarity stratum. The domestic-foreign differential ($\hat{\beta}_D - \hat{\beta}_F$) exhibits striking dynamics. For high-similarity stocks, the differential shifts from -0.013 in the pre-COVID period to $+0.634$ during COVID—a swing driven primarily by the foreign coefficient turning strongly negative ($\hat{\beta}_F = -0.578$) while the domestic coefficient remains slightly positive ($\hat{\beta}_D = +0.056$) during the pandemic quarter. The differential then reverses sign in the post-COVID era, returning toward its pre-pandemic level.

Panel B reports the triple-interaction coefficient $\hat{\beta}_7$ from:

$$\begin{aligned}\text{next_loss}_{i,t+1} = \beta_0 + \beta_1 s_{sd} + \beta_2 \text{domestic} + \beta_3 \text{post} + \beta_4 (\text{dom} \times \text{post}) \\ + \beta_5 (s_{sd} \times \text{dom}) + \beta_6 (s_{sd} \times \text{post}) + \beta_7 (s_{sd} \times \text{dom} \times \text{post}) + \varepsilon\end{aligned} \tag{11}$$

where *post* = 1 for COVID and post-COVID periods (2020Q1–2024Q4) and 0 for pre-COVID (2017Q1–2019Q4). A negative $\hat{\beta}_7$ indicates that the domestic sentiment advantage *strengthened* after COVID-19. The results show precisely this: $\hat{\beta}_7 = -0.035$ ($p = 0.074$) for low-similarity stocks and $\hat{\beta}_7 = -0.022$ ($p = 0.078$) for medium-similarity stocks, with an insignificant effect for high-similarity stocks ($+0.025$, $p = 0.428$).

Table XII: COVID-19 and the Domestic-Foreign Sentiment Gap

	$\hat{\beta}$ (s_sd → losses, QFE)			Differential ($\hat{\beta}_D - \hat{\beta}_F$)		
	High	Medium	Low	High	Medium	Low
Panel A: Domestic Predictive Coefficient by Period						
Pre-COVID (2017-2019)	+0.0034	+0.0591	+0.0731	−0.0128	−0.0413	+0.0944
COVID (2020)	+0.0556	+0.0206	−0.3617	+0.6339	+0.0921	+0.1715
Post-COVID (2021-2024)	−0.5633	−0.0763	−0.3661	−0.0153	−0.0988	+0.0889
Panel B: Triple Difference-in-Differences						
$\hat{\beta}_7$ (s_sd × domestic × post)	+0.0248		−0.0218*		−0.0352*	
	(0.0312)		(0.0124)		(0.0197)	

Notes: Panel A reports the predictive coefficient $\hat{\beta}$ from regressing next-quarter losses on sentiment volatility with quarter FE, estimated separately by period and stock type. The differential is $\hat{\beta}_D - \hat{\beta}_F$. Panel B reports the triple interaction coefficient from next_loss $= \beta_0 + \beta_1 s_{sd} + \beta_2 \text{domestic} + \beta_3 \text{post} + \beta_4(\text{dom} \times \text{post}) + \beta_5(s_{sd} \times \text{dom}) + \beta_6(s_{sd} \times \text{post}) + \beta_7(s_{sd} \times \text{dom} \times \text{post}) + \varepsilon$. A positive $\hat{\beta}_7$ indicates that the domestic sentiment advantage *weakened* after COVID. HC1 robust standard errors. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

This pattern has a precise economic interpretation. The negative triple-interaction coefficients for low- and medium-similarity stocks contradict a simple "physical proximity" story in which lockdowns equalize information access. Rather than disrupting local information advantages, the pandemic *amplified* them for stocks where information asymmetries were already most pronounced. During COVID-19, the volume and diversity of digital information expanded massively—Twitter's monetizable daily active users grew from 152 million (Q1 2019) to 217 million (Q4 2021)—but interpreting what this information meant for specific stocks required local context that domestic investors possessed.

The cross-stratum pattern is consistent with this mechanism. For high-similarity stocks, which are already well covered and informationally transparent, the within-COVID spike in the differential (+0.634, Panel A) proved temporary: the triple-interaction coefficient $\hat{\beta}_7 = +0.025$ ($p = 0.428$) captures the average domestic advantage over the entire 2020–2024 window relative to pre-COVID, and the large within-COVID spike is averaged with the post-COVID differential of −0.015, yielding an insignificant net effect consistent with a transient disruption rather than a persistent structural shift. For medium- and low-similarity stocks, where information asymmetries were already most pronounced, the domestic advantage measurably grew. This is consistent with a world where COVID increased the *quantity* of information available to all investors while simultaneously increasing the *complexity* of the information environment, raising the value of local contextual knowledge needed to interpret the flood of new signals.

We note two caveats. First, the "post" period (2021–2024) also saw the Elon Musk acquisition of Twitter (October 2022) and subsequent platform changes, which may have altered the composition of users and the information content of tweets. Our DID design attributes changes to the COVID shock, but late-sample dynamics could reflect platform-specific confounds. Second, the pre-COVID window is short (2017–2019, 12 quarters), limiting statistical power for the pre-period baseline. The insignificant $\hat{\beta}_7$ for high-similarity stocks could reflect either a true null or insufficient power to detect small effects in this stratum.

Figure IX visualizes the Q5–Q1 spread dynamics for domestic and foreign stocks across the three similarity strata, with the COVID period shaded. The figure shows the domestic-foreign gap widening during 2020 and, for low-similarity stocks, remaining elevated post-COVID. Importantly, the pre-COVID quarters (2017Q1–2019Q4) exhibit roughly parallel trends between domestic and foreign spread series across all three strata, supporting the identifying assumption that the domestic-foreign differential would have remained stable absent the COVID shock. The divergence begins precisely in the COVID window (yellow shading) rather than drifting gradually beforehand.

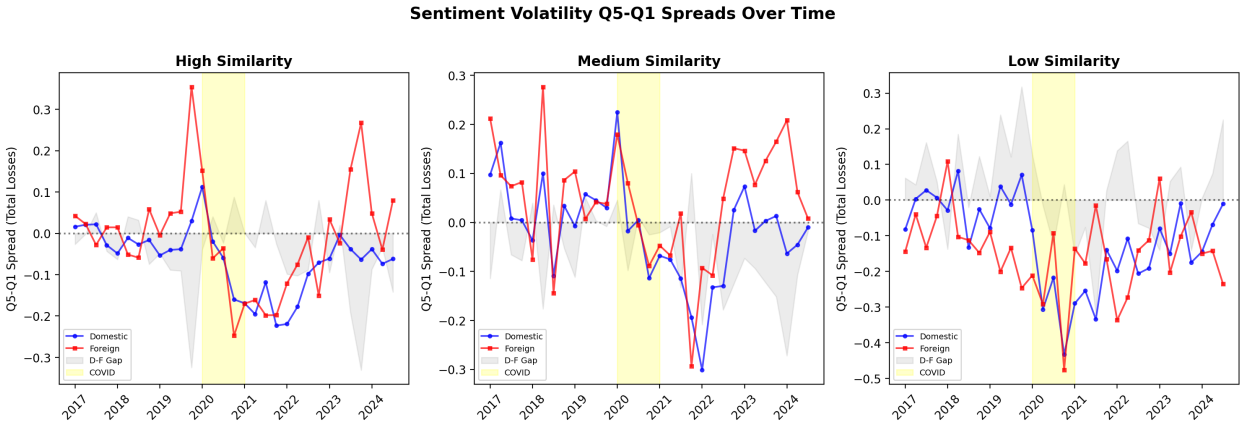


FIGURE IX
Sentiment Volatility Q5–Q1 Spreads Over Time

Notes: Quarterly Q5–Q1 spreads of next-quarter total losses sorted by lagged sentiment volatility (s_{sd}), plotted separately for domestic (blue circles) and foreign (red squares) stocks. Gray shading shows the domestic-foreign gap. Yellow shading indicates the COVID period (2020Q1–2020Q4). Each panel corresponds to a similarity stratum: high ($\phi \geq 0.98$), medium ($0.90 \leq \phi < 0.98$), and low ($\phi < 0.90$). Post-2017 sample; stratum-specific N as reported in Table XII ($N = 42{,}963$ high, $N = 47{,}570$ medium, $N = 20{,}992$ low).

7. Conclusion

This paper examines whether social media sentiment contains information about future stock losses and gains and, if so, for which stocks this information is most relevant. We propose an economic mechanism based on information frictions. U.S. Twitter sentiment is more informative for domestic stocks, where U.S.-based investors possess informational advantages about domestic firms, than for foreign stocks, where investors based in those companies' home countries hold superior information about firm fundamentals and domestic market conditions.

The empirical evidence supports this mechanism. For stocks that closely resemble the global cross-sectional benchmark, Twitter sentiment volatility reliably predicts larger subsequent losses for domestic stocks but not foreign stocks. This domestic advantage is strongest for stocks of intermediate benchmark similarity and collapses entirely for the least typical stocks, where domestic and foreign sentiment become equally informative. Sentiment disagreement tells a complementary story: for atypical stocks, high disagreement among Twitter users foreshadows compressed future gains, while for typical stocks the effect is absent or reversed. These patterns hold across two distinct measurement regimes and are not driven by aggregate macroeconomic conditions.

Structural estimation confirms that the domestic advantage reflects genuine signal precision rather than a mechanical consequence of domestic stocks being more volatile. For high-similarity stocks, domestic sentiment reliably predicts losses quarter after quarter regardless of macroeconomic conditions, while foreign sentiment oscillates in sign. For the least typical stocks, where the raw spreads converge, structural tests confirm that both investor groups achieve equivalent signal quality: the convergence reflects true parity in information content, not a statistical artifact.

Exploiting COVID-19 as a shock to the information landscape, we find that the pandemic amplified rather than equalized the domestic information advantage for the least typical and intermediate-similarity stocks. This result contradicts the simple story that physical proximity drives country-level information advantages: digital connectivity did not eliminate the domestic edge. As the volume of information exploded, so did the value of country-level investor knowledge needed to interpret it.

Several directions merit future research. First, examining the mechanism at higher frequency could reveal intraday dynamics of sentiment incorporation. Second, comparing U.S. sentiment

effects for ADRs with home-market sentiment effects for the underlying shares would provide direct evidence on relative information content. Third, extending the analysis to non-English social media platforms would test whether the information geography patterns generalize beyond U.S.-centered platforms.

The findings have implications for both academics and practitioners. For academics, the results demonstrate that reduced-form predictability of social media sentiment reflects economically meaningful information asymmetries rather than statistical artifacts. More broadly, our findings connect to the emerging literature on the "data economy" in finance: Farboodi et al. (2022) document that the value of data has diverged across firms as technology advanced, and our results show an analogous divergence across *signal sources*—crowd-sourced retail sentiment generates factor-adjusted abnormal returns for low-similarity stocks but is subsumed by known risk factors for high-similarity stocks. This geography-dependent value of data has implications for the design of alternative data products and for understanding why the explosion of social media information has not eliminated home bias.

For practitioners, the differential patterns across stock types suggest that sentiment-based strategies are most effective for low-similarity stocks. For atypical stocks, sentiment-based strategies generate economically large returns even after accounting for standard risk factors—on the order of 13 to 18 percent annualized. For typical stocks, no such excess return survives risk adjustment. Practitioners applying sentiment signals uniformly across their investment universe are leaving performance on the table. The COVID-era amplification of the domestic-foreign differential for low- and medium-similarity stocks further suggests that the value of country-level sentiment signals is not static but increases during periods of heightened information complexity.

References

Admati, Anat R., "A Noisy Rational Expectations Equilibrium for Multi-Asset Securities Markets," *Econometrica*, 1985, *53* (3), 629–657. 24

Antweiler, Werner and Murray Z. Frank, "Is All that Talk Just Noise? The Information Content of Internet Stock Message Boards," *Journal of Finance*, 2004, *59* (3), 1259–1294. 23

Barber, Brad M and Terrance Odean, "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors," *The Review of Financial Studies*, 2008, *21* (2), 785–818. 3, 23

Bartov, Eli, Lucile Faurel, and Partha S. Mohanram, "Can Twitter Help Predict Firm-Level Earnings and Stock Returns?," *The Accounting Review*, 2018, *93* (3), 25–57. 23

Chen, Hailiang, Prabuddha De, Yu Hu, and Byoung-Hyoun Hwang, "Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media," *Review of Financial Studies*, 2014, *27* (5), 1367–1403. 23

Cresci, Stefano, "A Decade of Social Bot Detection," *Communications of the ACM*, 2020, *63* (10), 72–83. 50

Fama, Eugene F. and Kenneth R. French, "Size, value, and momentum in international stock returns," *Journal of Financial Economics*, 2012, *105* (3), 457–472. 17

Farboodi, Maryam, Adrien Matray, Laura Veldkamp, and Venky Venkateswaran, "Where Has All the Data Gone?," *The Review of Financial Studies*, 2022, *35* (7), 3101–3138. 24, 25, 26, 27, 29, 33, 46

French, Kenneth R and James M Poterba, "Investor diversification and international equity markets," *American Economic Review*, 1991, *81* (2), 222–226. 1

Lettau, Martin and Sydney Ludvigson, "Consumption, Aggregate Wealth, and Expected Stock Returns," *The Journal of Finance*, 2001, *56* (3), 815–849. 20, 22, 49

Ludvigson, Sydney C., Sai Ma, and Serena Ng, "Uncertainty and Business Cycles: Exogenous Impulse or Endogenous Response?," *American Economic Journal: Macroeconomics*, 2021, *13* (4), 369–410. 20, 22, 27, 49

Merton, Robert C, "A simple model of capital market equilibrium with incomplete information," *The Journal of Finance*, 1987, *42* (3), 483–510. 3, 23

Mondria, Jordi and Thomas Wu, "The Puzzling Evolution of the Home Bias, Information Processing and Financial Openness," *Journal of Economic Dynamics and Control*, 2010, *34* (5), 875–896. 1, 23

Nieuwerburgh, Stijn Van and Laura Veldkamp, "Information immobility and the home bias puzzle," *The Journal of Finance*, 2009, *64* (3), 1187–1215. 1, 3, 23

Paydarzarnaghi, Mahnaz and David Rakowski, "Twitter, Investor Demographics, and the Diffusion of Information in Financial Markets," 2023. Working paper, University of Texas at Arlington. Presented at AFFI 2023. 1, 4

Peng, Lin and Wei Xiong, "Investor Attention, Overconfidence and Category Learning," *Journal of Financial Economics*, 2006, *80* (3), 563–602. 23

Shleifer, Andrei and Robert W Vishny, "The limits of arbitrage," *The Journal of Finance*, 1997, *52* (1), 35–55. 2, 23

Sims, Christopher A., "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690. 1, 3, 23

Tetlock, Paul C., "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *Journal of Finance*, 2007, *62* (3), 1139–1168. 23

Twitter, Inc., "Annual Report 2019," 2019. Form 10-K filed with the U.S. Securities and Exchange Commission. U.S. monetizable daily active users: 31 million of 152 million global mDAU. Available at https://www.annualreports.com/Company/twitter-inc. 1, 4

Appendix A. Data Construction Details

This appendix provides detailed descriptions of data sources, variable construction, and preprocessing procedures.

Appendix A.1. Sample Construction

The analysis combines three data sources.

CRSP-Compustat Merged Panel. We begin with 471,879 firm-quarter observations across 17,302 unique firms from 2012Q1–2024Q4. The panel includes fundamentals, trading activity, factor exposures, and market conditions.

Context Analytics Twitter Sentiment. Daily sentiment data are aggregated to quarterly frequency. Context Analytics provides the Raw-S metric, measuring the sum of sentiment scores from tweets mentioning each stock. We construct three measures.

- `s_sd`: Standard deviation of daily Raw-S within each quarter
- `s_dispersion_mean`: Quarterly mean of daily S_DISPERSION (source diversity metric)
- `signal_vol_sum`: Sum of daily sentiment $\times$ log(1 + tweet volume)

Macroeconomic Controls. Consumption-wealth-income trend deviation (CAY) from Lettau and Ludvigson (2001) and financial uncertainty index from Ludvigson et al. (2021).

Appendix A.2. Similarity Index Construction

We measure how closely each stock's characteristics resemble the typical stock in our sample using a two-stage procedure.

Stage 1. Feature Importance. We estimate a Random Forest model predicting quarterly excess returns using 72 predetermined firm characteristics. Features include size measures, Fama-French factor betas, lagged volatility, liquidity proxies, and fundamentals. Out-of-bag R^2 is 4.28%.

Stage 2. Weighted Cosine Similarity. For each stock i in quarter t, we compute the weighted cosine similarity

$$\text{similarity}_{i,t} = \frac{\sum_{k=1}^{72} w_k \, x_{i,t,k} \, b_{t,k}}{\sqrt{\sum_{k=1}^{72} w_k \, x_{i,t,k}^2} \sqrt{\sum_{k=1}^{72} w_k \, b_{t,k}^2}} \tag{A.1}$$

where $w_k \geq 0$ are the normalized Gini-based Random Forest importance weights ($\sum_{k=1}^{72} w_k = 1$), $x_{i,t,k}$ is stock i's standardized feature k, and $b_{t,k}$ is the median of feature k across *all stocks*, both domestic and foreign, in quarter t.

The resulting distribution has mean 0.90, median 0.97, indicating most stocks closely resemble the global benchmark. See Figure I for the full distribution.

Appendix A.3. Context Analytics Methodology Change and Platform Evolution

In 2017, Context Analytics transitioned from cashtag-only tweet identification ($AAPL) to a topic model approach using company names and abbreviations. This produced an 85% reduction in sentiment levels. We use post-2017 as the primary sample and verify robustness using pre-2017 data.

The 2017 methodology change coincided with broader shifts in the Twitter platform that affect the interpretation of our sentiment measures. First, Twitter's user base grew substantially during our sample period: monetizable daily active users (mDAU) increased from approximately 109 million in Q1 2017 to 229 million by Q1 2022, more than doubling the volume of potential sentiment signals. This growth implies that the information environment from which Context Analytics extracts sentiment changed in both scale and composition over the sample period.

Second, the prevalence and sophistication of automated accounts ("bots") on Twitter evolved markedly. Research on bot detection grew rapidly during our sample period, with the number of published studies increasing from approximately 74 in 2015 to 214 by 2019 (Cresci, 2020). This literature documents that bot activity on social media platforms is substantial and potentially capable of distorting aggregate sentiment measures. Context Analytics applies proprietary filtering to exclude automated accounts, but the effectiveness of such filtering against increasingly sophisticated bots is uncertain.

These platform-level changes motivate two design choices. First, our primary analysis uses the post-2017 sample, which employs a consistent methodology throughout and covers the period of most rapid user growth. Second, we report pre-2017 robustness checks (Section 6.2) to verify that the core findings are not artifacts of either the methodology change or the evolving user composition. The consistency of sign and magnitude for sentiment dispersion effects across both periods (Figure VIII) provides reassurance that the documented information geography patterns are not driven by platform-specific confounds.

Appendix A.4. Variable Definitions

Table A.13: Variable Definitions

Variable	Definition
`s_sd`	Std. dev. of daily Raw-S sentiment within quarter
`s_dispersion_mean`	Mean of daily S_DISPERSION (source diversity, 0–1)
`total_loss_excess_ret`	Within-quarter sum of $\min(r^e_{id}, 0)$ across trading days; weakly negative by construction
`total_gain_excess_ret`	Within-quarter sum of $\max(r^e_{id}, 0)$ across trading days; weakly positive by construction
`deviation_quarterly`	Within-quarter sum of daily excess returns minus within-quarter sum of daily S&P 500 returns (`sprtrn`)
`sd_excess_ret`	Within-quarter standard deviation of daily excess returns
`similarity_index`	Weighted cosine similarity to global benchmark (all stocks)
`foreign`	Indicator: 1 if foreign company (trading via ADR), 0 if domestic